

2018 Machine Learning with R

Association Rules

강 필 성

고려대학교 산업경영공학부

pilsung_kang@korea.ac.kr

목차

I

연관규칙분석: A Priori Algorithm

II

연관규칙분석: 활용 분야

III

R 실습

한번쯤은 봤을 법한...



크게 보기 | 미리 보기

영업점 재고 · 위치 >

Klover 평점 102명

안내

오늘의책 북모닝 CEO 이벤트 무료배송 사은품

호모 데우스 미래의 역사

유발 하라리 지음 | 김명주 옮김 | 김영사 | 2017년 05월 19일 출간

★★★★★ 리뷰 30개 [리뷰쓰기](#)

국내도서 주간베스트 8위 | 인문 주간베스트 2위

주요 일간지 북섹션 추천도서 ▾

정가 : 22,000원

판매가 : **19,800원** [10%↓ 2,200원 할인]

제휴할인가 : **18,810원** 교보-KB국민카드 5% 청구할인(실적무관) [카드/포인트 안내](#)

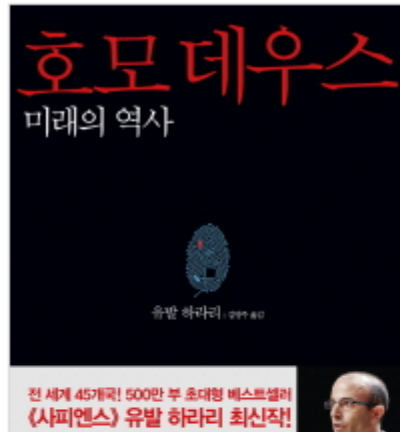
통합포인트 : **1,100원** 적립 [5% 적립]

추가혜택 : **N Pay** 네이버페이 결제 시 최대 **2%** 추가 적립 >

PAYCO 페이코 결제 시 **6,500원** 할인 + **1만원** 적립 >

L.POINT 실 결제 금액의 **0.5%** 적립 [안내](#)

한번쯤은 봤을 법한...



오늘의책 북모닝CEO 이벤트 무료배송 사은품

호모 데우스 미래의 역사

유발 하라리 지음 | 김명주 옮김 | 김영사 | 2017년 05월 19일 출간

★★★★★ 리뷰 30개 [리뷰쓰기](#)

국내도서 주간베스트 8위 | 인문 주간베스트 2위

주요 일간지 북섹션 추천도서 ▼

정가 : 22,000원

이 책을 구매하신 분들이 함께 구매하신 상품입니다

전체선택

장바구니 담기



말의 품격

13,050원



라틴어 수업(지적이고 아름다운 삶을 위한)

13,500원



인간의 위대한 여정(양장본 HardCover)

19,800원



사피엔스

19,800원



이탈리아의 사생활

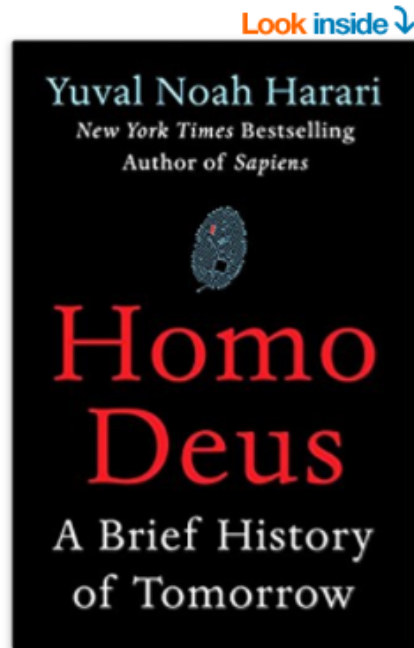
14,400원



한식의 품격(반양장)

16,200원

한번쯤은 봤을 법한...



 Listen



See this image

Homo Deus: A Brief History of Tomorrow Hardcover – February 21, 2017

by Yuval Noah Harari (Author)

★★★★☆ 444 customer reviews

Amazon Charts #16 Most Read

▶ See all 14 formats and editions

Kindle
\$15.50


Read with Our **Free App**

Hardcover
\$19.88

12 Used from \$15.00
51 New from \$15.25
1 Collectible from \$84.95

Paperback
from \$8.99

15 Used from \$9.48
34 New from \$8.99

 Audiobook
\$0.00

Free with your Audible trial

Audio CD
\$30.29

5 Used from \$24.17
27 New from \$21.57

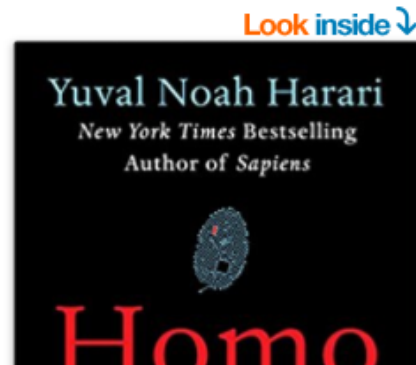
NEW YORK TIMES BESTSELLER

Yuval Noah Harari, author of the critically-acclaimed *New York Times* bestseller and international phenomenon *Sapiens*, returns with an equally original, compelling, and provocative book, turning his focus toward humanity's future, and our quest to upgrade humans into gods.

Over the past century humankind has managed to do the impossible and rein in famine, plague, and war. This may seem hard to accept, but, as Harari explains in his trademark style—thorough, yet riveting—famine, plague and war have been transformed from incomprehensible and uncontrollable forces of nature into manageable challenges. For the first time ever, more people die from eating too much than

▶ Read more

한번쯤은 봤을 법한...



Look inside ↴

Homo Deus: A Brief History of Tomorrow Hardcover – February 21, 2017

by Yuval Noah Harari (Author)

★★★★☆ 444 customer reviews

Amazon Charts #16 Most Read

See all 14 formats and editions

Kindle
\$15.50

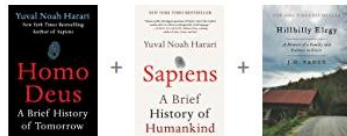
Hardcover
\$19.88

Paperback
from \$8.99

Audiobook
\$0.00

Audio CD
\$30.29

Frequently bought together



Total price: **\$57.67**

Add all three to Cart

Add all three to List

✓ **This item:** Homo Deus: A Brief History of Tomorrow by Yuval Noah Harari Hardcover **\$19.88**

✓ **Sapiens:** A Brief History of Humankind by Yuval Noah Harari Hardcover **\$21.00**

✓ **Hillbilly Elegy:** A Memoir of a Family and Culture in Crisis by J. D. Vance Hardcover **\$16.79**

Customers who bought this item also bought

Page 1 of 13

Sapiens: A Brief History of Humankind
Yuval Noah Harari
★★★★☆ 3,286
Hardcover
\$21.00 ✓prime

A Full Life: Reflections at Ninety
Jimmy Carter
★★★★☆ 261
Hardcover
64 offers from **\$7.15**

The Inevitable: Understanding the 12 Technological Forces...
Kevin Kelly
★★★★☆ 225
Paperback
\$12.18 ✓prime

The Heart: A Novel
Maylis de Kerangal
★★★★☆ 82
Hardcover
\$19.93 ✓prime

The Heart: A Novel
Maylis de Kerangal
★★★★☆ 82
Paperback
\$11.00 ✓prime

Hillbilly Elegy: A Memoir of a Family and Culture in Crisis
J. D. Vance
★★★★☆ 7,255
Hardcover
\$16.79 ✓prime

Summary: Sapiens: A Brief History of Humankind
Yuval Noah Harari
★★★★☆ 14
Paperback
\$9.99 ✓prime

Easternization: Asia's Rise and America's Decline from Obama to Trump...
Gideon Rachman
★★★★☆ 12
Hardcover
\$17.07 ✓prime

일상 상황에서의 연관 규칙 분석

❖ 장바구니 분석: Market basket analysis (MBA)



Wall Mart (USA)



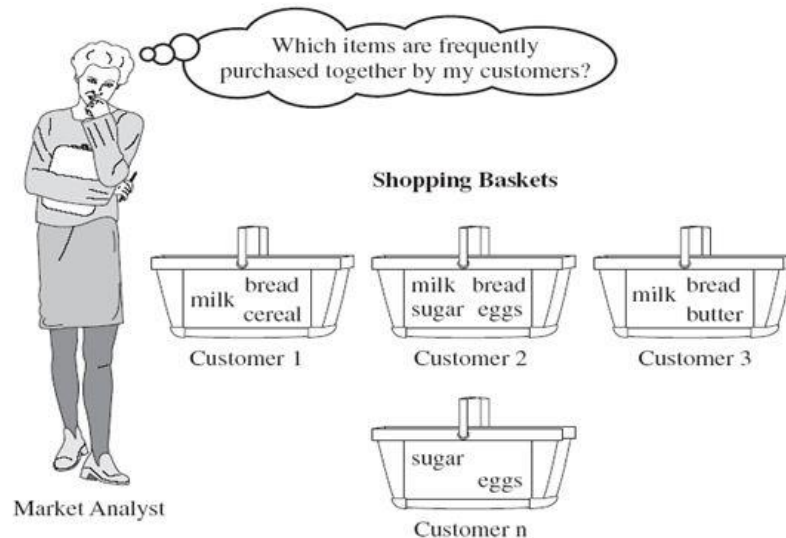
E-Mart (Korea)



연관규칙 분석

❖ 목적

- 어떤 두 아이템 집합이 **빈번히 발생**하는가를 알려주는 **일련의 규칙**들을 생성
 - ✓ Produce rules that define “what goes with what”
- 우리의 데이터에 의하면 “X 아이템을 구매하는 고객들은 Y 아이템 역시 구매할 가능성이 높다” → 콘텐츠 기반 추천 시스템에 널리 사용
- 장바구니 분석(Market Basket Analysis)으로도 널리 알려짐



연관규칙 분석

❖ 데이터 속성









- 각 레코드는 트랜잭션의 형태를 가짐
- 행렬의 형태로 표현하게 되면 대부분의 셀이 0의 값은 갖는 희소행렬(sparse matrix)이 됨

Tid	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Tid	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

연관규칙 분석: 예제

❖ 동네 작은 가게 매출 트랜잭션 데이터

Transaction	Item 1	Item 2	Item 3	Item 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

연관규칙 분석: 용어 및 규칙 생성

❖ 용어: Terminology

- 조건절(Antecedent) – “IF” part
- 결과절(Consequent) – “THEN” part
- 아이템 집합(Item set) – 조건절 또는 결과절을 구성하는 아이템들의 집합
- 조건절 아이템 집합과 결과절 아이템 집합은 상호배반 (한 아이템이 조건절과 결과절에 모두 포함될 수 없음)

❖ 규칙 생성: Generating rules

- 매우 많은 수의 규칙이 생성 가능 (예시: 첫번째 트랜잭션)
 - ✓ 계란을 구매하는 사람들은 라면도 함께 구매한다.
 - ✓ 계란과 라면을 구매하는 사람들은 참치도 함께 구매한다.
 - ✓ 참치를 구매하는 사람들은 계란도 함께 구매한다.
 - ✓ ...

연관규칙 분석: 규칙의 효용성 측정 지표

For the rule $A \rightarrow B$

❖ 지지도: Support

$$\text{support}(A \rightarrow B) = P(A) \text{ or } P(A, B)$$

- 빈발 아이템 집합(frequent item sets)을 판별하는데 사용

❖ 신뢰도: Confidence

$$\text{confidence}(A \rightarrow B) = \frac{P(A, B)}{P(A)}$$

- 아이템 집합 간의 연관성 강도를 측정하는데 사용

❖ 향상도: Lift

$$\text{lift}(A \rightarrow B) = \frac{P(A, B)}{P(A) \cdot P(B)}$$

- 생성된 규칙이 실제 효용가치가 있는지를 판별하는데 사용

연관규칙 분석: 규칙의 효용성 측정 지표

Rule: $X \Rightarrow Y$

$$Support = \frac{freq(X, Y)}{N}$$

$$Confidence = \frac{freq(X, Y)}{freq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$



Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

연관규칙 분석: 규칙 생성

❖ 유용한 연관 규칙들을 어떻게 찾아낼 것인가?

- 이상적으로는 모든 생성 가능한 규칙을 만든 뒤, 각 규칙의 지지도, 신뢰도, 향상도를 측정하여 유용한 규칙들만을 찾아냄
- 아이템 수가 증가할수록 계산에 소요되는 시간이 기하급수적으로 증가함

❖ Brute-force approach

- 가능한 모든 규칙을 나열함
- 모든 규칙의 지지도와 신뢰도를 계산함
- 최소지지도와 최소신뢰도 조건을 만족하지 못하는 규칙을 제거
- **Computationally prohibitive!**

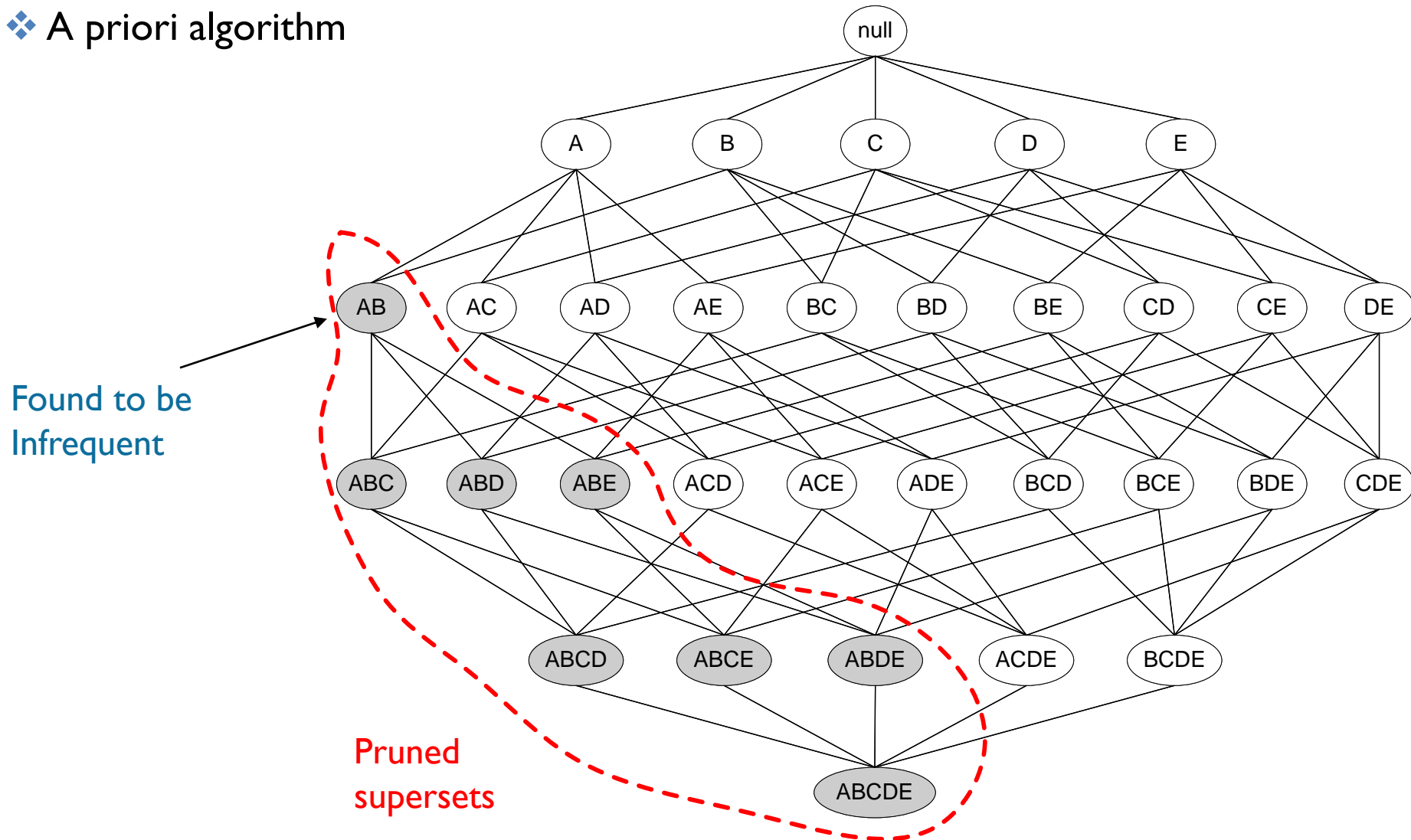
연관규칙 분석: 규칙 생성

❖ A priori algorithm

- 빈발 집합(frequent item sets)만을 고려하여 규칙 생성
- **지지도(support)**
 - ✓ 조건절에 속하는 아이템 집합이 발생할 확률
 - ✓ 아이템 집합 {계란, 라면}의 지지도는 40%
- **최소 지지도(minimum support)**
 - ✓ 유용한 규칙으로 인정받기 위해 필요한 최소 지지도
- 최소 지지를 만족하지 못하는 아이템 집합의 상위집합(superset)은 항상 최소 지지를 만족하지 않음
 - ✓ Support of an item set never exceeds the support of its subsets, which is known as **anti-monotone** property of support.

연관규칙 분석: 규칙 생성

❖ A priori algorithm



연관규칙 분석: 빈발 아이템 집합 생성

최소 지지도 조건 부여

- 최소 지지도: 2 transactions or 20%

Transaction	Item 1	Item 2	Item 3	Item 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

연관규칙 분석: 빈발 아이템 집합 생성

2

최소 지지도 조건을 만족하는 1개짜리 아이템 집합을 생성

- $\text{Support \{noodle\}} = 8/10 = 80\%$
- $\text{Support \{egg\}} = 5/10 = 50\%$
- $\text{Support \{cola\}} = 5/10 = 50\%$
- $\text{Support \{rice\}} = 3/10 = 30\%$
- $\text{Support \{tuna\}} = 2/10 = 20\%$
- $\text{Support \{onion\}} = 1/10 = 10\%$

양파(onion)는 최소 지지도 조건을 만족하지 못했으므로 이후 분석에서 제외

연관규칙 분석: 빈발 아이템 집합 생성

3

앞 단계에서 살아남은 아이템들을 이용하여 최소 지지도 조건을 만족하는 2개짜리 아이템 집합을 생성

	noodle	egg	cola	rice	tuna
noodle		40%	40%	20%	20%
egg			30%	0%	20%
cola				0%	10%
rice					0%
tuna					

- {noodle, egg}, {noodle, cola}, {noodle, rice}, {noodle, tuna}, {egg, cola}, {egg, tuna} are frequent two-item sets.

연관규칙 분석: 빈발 아이템 집합 생성

더 이상 최소 지지도 이상을 나타내는 아이템 집합이 없을 때까지 아
이템 집합의 크기를 1씩 증가시키면서 반복 수행

4

Set-size	Item 1	Item 2	Item 3	...	Item 6
1	noodle				
1	egg				
1	cola				
1	rice				
1	tuna				
2	noodle	egg			
2	noodle	cola			
2	noodle	rice			
...			

연관규칙 분석: A Priori Algorithm

❖ A priori algorithm

- Let $k=1$
- Generate frequent itemsets of length 1
- Repeat until no new frequent itemsets are identified
 - ✓ Generate length $(k+1)$ candidate itemsets from length k frequent itemsets
 - ✓ Prune candidate itemsets containing subsets of length k that are infrequent
 - ✓ Count the support of each candidate by scanning the DB
 - ✓ Eliminate candidates that are infrequent, leaving only those that are frequent

연관규칙 분석: 규칙 평가

❖ 신뢰도: Confidence

- 조건절이 발생했다는 가정 하에 결과절이 발생할 조건부 확률
- E.g. “if noodle is purchased, then egg is also purchased”

$$\text{support}(\text{noodle}) = P(\text{noodle}) = \frac{8}{10}, \quad \text{support}(\text{egg}) = P(\text{egg}) = \frac{5}{10}$$

$$\text{confidence}(\text{noodle} \rightarrow \text{egg}) = \frac{P(\text{noodle}, \text{egg})}{P(\text{noodle})} = \frac{4/10}{8/10} = 0.5(50\%)$$

- 비교 대상 신뢰도(benchmark confidence): 전체 트랜잭션에서 결과절이 발생할 확률 ($P(\text{egg})$, $\text{support}(\text{egg})$)
- 규칙 ($\text{noodle} \rightarrow \text{egg}$)의 신뢰도가 $P(\text{egg})$ 보다 작으면 규칙으로서의 효용 가치는 낮음

연관규칙 분석: 규칙 평가

❖ 지지도: Lift

- 신뢰도/비교 대상 신뢰도: Confidence/(benchmark confidence)

$lift(noodle \rightarrow egg)$

$$\begin{aligned}
 &= \frac{confidence(noodle \rightarrow egg)}{support(egg)} = \frac{P(noodle, egg)}{P(egg)} = \frac{P(noodle, egg)}{P(noodle) \times P(egg)} \\
 &= \frac{\frac{4}{10}}{\frac{8}{10} \times \frac{5}{10}} = 1
 \end{aligned}$$

- 신뢰도 = 1일 경우, 조건절과 결과절은 통계적으로 독립사건임을 의미함 ➔ 규칙 사이에 유의미한 연관성이 없음
- 신뢰도 > 1일 경우 조건절과 결과절은 서로 긍정적인 연관관계를 나타냄

연관규칙 분석: 사례 결과

❖ 규칙 생성을 위한 기준 지지도 및 신뢰도 설정

- 기준 지지도: 20%
- 기준 신뢰도: 70%

Rule #	Antecedent (a)	Consequent	Support	Confidence	Lift
1	tuna=>	egg, noodle	2	100	2.5
2	tuna=>	egg	2	100	2
3	noodle, tuna=>	egg	2	100	2
4	rice=>	noodle	3	100	1.25
5	egg, tuna=>	noodle	2	100	1.25
6	tuna=>	noodle	2	100	1.25
7	cola=>	noodle	5	80	1
8	egg=>	noodle	5	80	1

연관규칙 분석: 요약

❖ 연관규칙분석

- 트랜잭션 데이터베이스에 존재하는 아이템 집합들 간의 연관성을 나타내는 규칙을 생성하는 분석 기법
- 다양한 분야의 추천 시스템 구축에 널리 사용됨
- 전체 규칙을 모두 생성하는 것이 비효율적이기 때문에 효율적인 빈발 집합을 찾아내는 A Priori 알고리즘을 사용
- 규칙의 효용성은 지지도, 신뢰도, 향상도의 세 가지를 이용하여 평가
- 규칙 1: $A \rightarrow B$ 와 규칙 2: $C \rightarrow D$ 에 대해 지지도, 신뢰도, 향상도가 모두 클 경우에만 규칙 1이 규칙 2보다 효과적인 규칙으로 결론지을 수 있음

순차연관분석: Sequential association rule mining

❖ 질문: 시간 순서를 고려한 순차적 연관성을 판단할 수 있는가?

■ 순차연관분석에서 사용하는 데이터셋

- ✓ 시간에 대한 정보(Timestamp)와 사용자 또는 이벤트(행위의 주체) 정보가 함께 포함되어야 함
- ✓ 데이터셋 예시 (R의 “arulesSequence” 패키지에서 사용하는 형식)

Itemset	sequenceID	eventID	SIZE
{A,B}	1	1	2
{C,D,E}	1	2	3
{A,D,F}	1	3	3
...
{C,F,G}	2	1	2
{F,G,H,I}	2	2	4
...

1번 고객이 첫 번째 구매 시 A,B를 구매

1번 고객이 두 번째 구매 시 C,D,E를 구매

1번 고객이 세 번째 구매 시 A,D,F를 구매

2번 고객이 첫 번째 구매 시 C,F,G를 구매

2번 고객이 두 번째 구매 시 F,G,H,I를 구매

순차연관분석: Sequential association rule mining

❖ 순차연관분석

■ Subsequence

✓ 한 sequence ID 에 등장하는 item의 순서가 다른 sequence ID에 등장하는 item의 순서의 일부분일 때

- $B \rightarrow A$ 는 $AB \rightarrow E \rightarrow ACD$ 의 subsequence
- $AB \rightarrow E$ 는 ABE 의 subsequence가 아님 (순서가 보존되지 않음)

순차연관분석: Sequential association rule mining

❖ 순차연관분석

■ Subsequence

✓ Sequence DB 예시 및 Frequent sequence

DATABASE			FREQUENT SEQUENCES	
SID	Time (EID)	Items	Frequent 1-Sequences	
1	10	C D	A	4
1	15	A B C	B	4
1	20	A B F	D	2
1	25	A C D F	F	4
2	15	A B F	Frequent 2-Sequences	
2	20	E	AB	3
3	10	A B F	AF	3
4	10	D G H	B→A	2
4	20	B F	BF	4
4	25	A G H	D→A	2
			D→B	2
			D→F	2
			F→A	2
			Frequent 3-Sequences	
			ABF	3
			BF→A	2
			D→BF	2
			D→B→A	2
			D→F→A	2
			Frequent 4-Sequences	
			D→BF→A	2

순차연관분석: Sequential association rule mining

❖ 순차연관분석

■ 기본 절차

- ✓ 1단계: 최소 지지도(minimum support)를 만족하는 모든 sequence들의 집합 A 를 찾는다
- ✓ 2단계: A 에 속하는 개별 sequence β 들에 대해 다음 과정 수행
- ✓ 2-1단계: 해당 sequence에서 시간 순으로 앞에 위치한 subsequence α_{before} 와 뒤에 위치한 subsequence α_{after} 대해 confidence를 계산

$$conf = fr(\alpha_{after} \cup \alpha_{before}) / fr(\alpha_{before})$$

- ✓ 2-2단계: 최소 confidence 값을 넘어서는 subsequence들에 대해 최종 규칙 생성

$$\alpha_{before} \rightarrow \alpha_{after}$$

- 대량의 데이터베이스에서 효율적으로 frequent sequence를 찾아가는 여러 알고리즘이 존재: GSP algorithm, Sequential PAttern Discovery using equivalence classes (SPADE) 등

목차

I

연관규칙분석: A Priori Algorithm

II

연관규칙분석: 활용 분야

III

R 실습

사례 1: 유통에서의 연관규칙분석

❖ 국내 실제 유통회사 데이터를 이용한 분석 사례

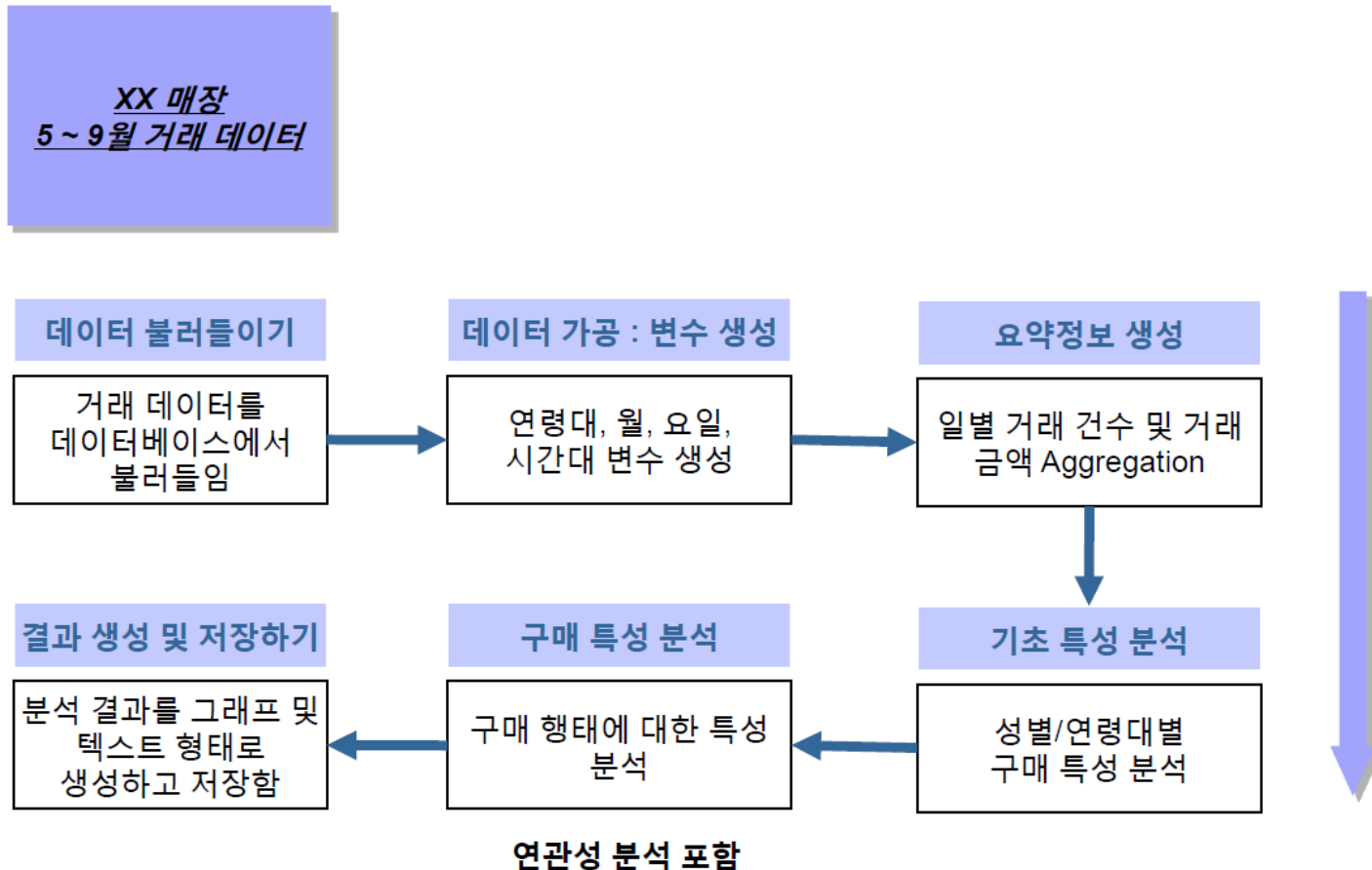
■ 데이터 구조

변수명	설명
회원번호	개인 고유 식별 번호
성별	남자 or 여자
연령대	회원의 생년월일을 기준으로 연령대 구분
발급일자	회원 가입일자
매출일자	상품 구매 일자
매출요일	상품 구매 요일
평일공휴일구분	평일 혹은 공휴일 구분 (공휴일은 토요일, 일요일, 휴일)
대분류명	상품의 대분류 구분
중분류명	상품의 중분류 구분
소분류명	상품의 소분류 구분
세분류명	상품의 세분류 구분
수량	상품 구매 개수
매출금액	상품 구매 금액

사례 1: 유통에서의 연관규칙분석

❖ 국내 실제 유통회사 데이터를 이용한 분석 사례

■ 데이터 분석 프로세스



사례 1: 유통에서의 연관규칙분석

❖ 국내 실제 유통회사 데이터를 이용한 분석 사례

■ 데이터 분석 시나리오

	<u>시나리오 명</u>	<u>분석 방향</u>
시나리오 1	기초 특성 분석	<ul style="list-style-type: none"> • 상세 분석을 위한 데이터 파악 • 구매 현황 리포트 작성
시나리오 2	그룹별 구매 패턴 분석	<ul style="list-style-type: none"> • 전체, 성별/연령대별 상품 구매 패턴 파악
시나리오 3	개인별 구매 패턴 분석	<ul style="list-style-type: none"> • 개인별 상품 구매 패턴 파악
시나리오 4	연관성 분석	<ul style="list-style-type: none"> • 주로 구매하는 상품으로 장바구니를 구성함

사례 1: 유통에서의 연관규칙분석

❖ 국내 실제 유통회사 데이터를 이용한 분석 사례

■ 구매 물품의 시각화



■ 연관규칙 분석을 통한 상품집합 도출

[illegible]

Graph for 12 rules

탁주

일반커피믹스

병맥주

페트콜라

일반생수

병소주

탄산과즙

페트맥주

천연과즙

페트사이다

사례 2: 동영상 클립 추천

❖ 광고 수익 증대를 위한 동영상 추천 시스템 구축

하루 평균 1억 건 이상의 데이터를 처리

동영상·광고의 로그 및 메타정보를 raw데이터로 저장·관리하며
수집 항목 및 사이트는 꾸준히 증가하고 있음

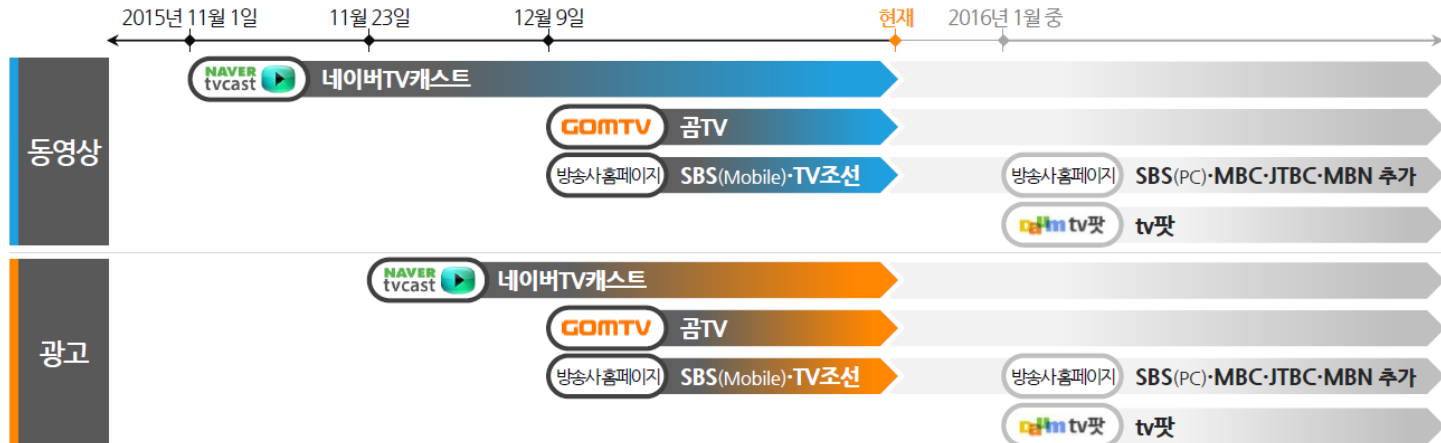
■ 하루 평균 처리하는 Data량 (15년 12월 기준)

1억 건(30억 건/월)

- ※ 월 평균 광고노출수 5억 * 광고로그 5개 + 동영상 조회수 5억 * 동영상로그 1개 + α (메타정보)
- ※ 현재 동영상로그가 1개이지만, 협의를 통해 이용시간 확보를 위해 5개로 늘릴 예정임
- ※ 데이터 신뢰성, 타당도 높음

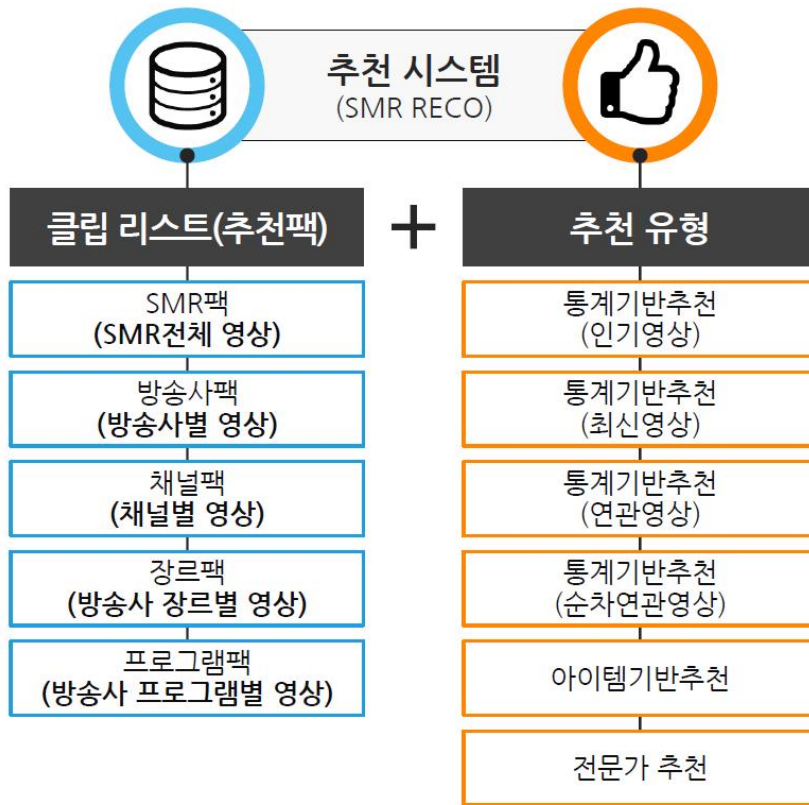
더보기

■ 데이터 수집 사이트 현황 (15년 12월 기준)



사례 2: 동영상 클립 추천


❖ 광고 수익 증대를 위한 동영상 추천 시스템 구축



사례 2: 동영상 클립 추천

❖ 광고 수익 증대를 위한 동영상 추천 시스템 구축

유희열의 스케치북 + 구독 17,185 방송 다시보기 구독 66,595명



음악이 필요한 순간 Melon 야생화 박효신
작사·박효신, 김지향 / 작곡·박효신, 정재일

▶ 1,336,050 | 등록 2016.10.30. | 3,152 | ^ 접기

" 아무리 불러도 무너지지 않는 곡 " 박효신 - 야생화 17,428

사례 2: 동영상 클립 추천

❖ 광고 수익 증대를 위한 동영상 추천 시스템 구축

함께 재생된 동영상



손승연, 전율의 퍼블보컬 '금지된 사랑'
불후의 명곡2 전설을 노래하다
▷ 243,551 ♡ 825



이승철&정인, EDM으로 재탄생 '서쪽 하늘'
불후의 명곡2 전설을 노래하다
▷ 226,822 ♡ 838



첫눈처럼 너에게 가겠다 - 에일리
유희열의 스케치북
▷ 711,441 ♡ 5,741



손승연, 파워풀한 가창력+애절한 '사랑 안 해'
불후의 명곡2 전설을 노래하다
▷ 266,040 ♡ 1,056



에일리, 압도적인 가창력 '편지'
불후의 명곡2 전설을 노래하다
▷ 363,955 ♡ 1,956



황치열&경리, '이브의 경고' 커플 댄스
불후의 명곡2 전설을 노래하다
▷ 24,410 ♡ 751



치열이의 치열한 신고식~! '복고 댄스'
불후의 명곡2 전설을 노래하다
▷ 9,623 ♡ 365



[1회] 지리산 소년 김영근 - "Lay Me Down"
슈퍼스타K 2016
▷ 1,965,325 ♡ 9,368



마마무, 센스 있는 개사가 돋보인 축하무대 '데칼코마니'
청룡영화상
▷ 2,795,060 ♡ 21,404



[1회] 한 곡으로는 모자라! 김영근 - "탈진"
슈퍼스타K 2016
▷ 1,897,812 ♡ 8,224



'아기해마' VS '다이빙 소년'의 1라운드 무대! - Sea of Love
복면가왕
▷ 88,800 ♡ 499



한동근&최효인, '거짓말 거짓말 거짓말' 역대급 소름 무대!~
듀엣가요제
▷ 494,583 ♡ 2,651



사랑스러운 목소리의 '올리브스녀'는 에이핑크 오하영!
복면가왕
▷ 139,772 ♡ 1,581



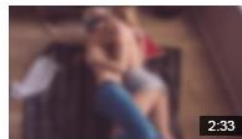
'불광동 휘발유'의 3라운드 무대! - 헤어지는 중입니다
복면가왕
▷ 555,697 ♡ 3,786



[15화 예고] 김고은, 공유 향한 작별인사?! '또 만나요.'
tvN 10주년 특별기획 <도깨비>
▷ 1,970,210 ♡ 3,149



영화 '리얼' 설리(최진리) 전라신 편집 없이 개봉 결정
뉴스컬처 연예TV
▷ 1,454,008 ♡ 937



영화 '리얼' 설리, 전라 노출 상영 인포케이
▷ 427,126 ♡ 321



[풀버전] 주노플로 @ 2차 예선 full ver.
소미더머니6
▷ 272,308 ♡ 2,603



황광희 X 개코 - 당신의 밤 (feat.오드무대! - 숲소리)
무한도전
▷ 2,386,613 ♡ 22,395



'올리브스녀' VS '뽀빠이'의 1라운드 무대! - 숲소리
복면가왕
▷ 69,793 ♡ 668



[유립피안 챔피언십] 이보다 더 완벽할 순 없다. 딕 야스퍼스
당구전문 인터넷 방송국 코즘코리아
▷ 15,197 ♡ 8

- 연구 목적: 근접무선통신기술(NFC)을 이용하여 수집한 박람회 관람객 행동 데이터를 분석하여 전시부스간 연관관계 분석 및 부스 배치 최적화
- 데이터 수집 체계



사례 3: 박람회 부스 이용 패턴 분석

❖ “2013 내나라 여행 박람회” 데이터를 이용한 분석 사례

- 연구 목적: 근접무선통신기술(NFC)을 이용하여 수집한 박람회 관람객 행동 데이터를 분석하여 전시부스 간 연관관계 분석 및 부스 배치 최적화



문화재청


통합검색 ▼ 검색어를 입력하세요
 

 국가상징 알아보기
  직원찾기
  주요홈페이지
  전체메뉴

새소식 | 국민참여 | 행정정보 | 정보공개 | 문화재청소개

새소식 | 입찰정보



새소식

- 공지사항
- 보도/해명
- 사진뉴스
- 공감! 문화재
- 문화재 지정예고
- 입찰정보**
- 시험/채용
- 문화유산 유유자적
- 문화재청 소식지 +

입찰정보





입찰공고번호	2017-3	구분	입찰정보
건명	2017 내나라 여행박람회 문화재청 전시부스 설치 운영 용역		
입찰일자	입찰공고서 참조	게시일	20170104103321
발주부서	활용정책과	담당자	임미정

- 용역명 : 2017 내나라 여행박람회 문화재청 전시부스 설치 운영 용역
- 예산액 : 금48,800,000원
- 세부내용 : 입찰공고서 등 첨부물 참조
- 사업담당 : 활용정책과 형은희(042-481-4747)

첨부파일

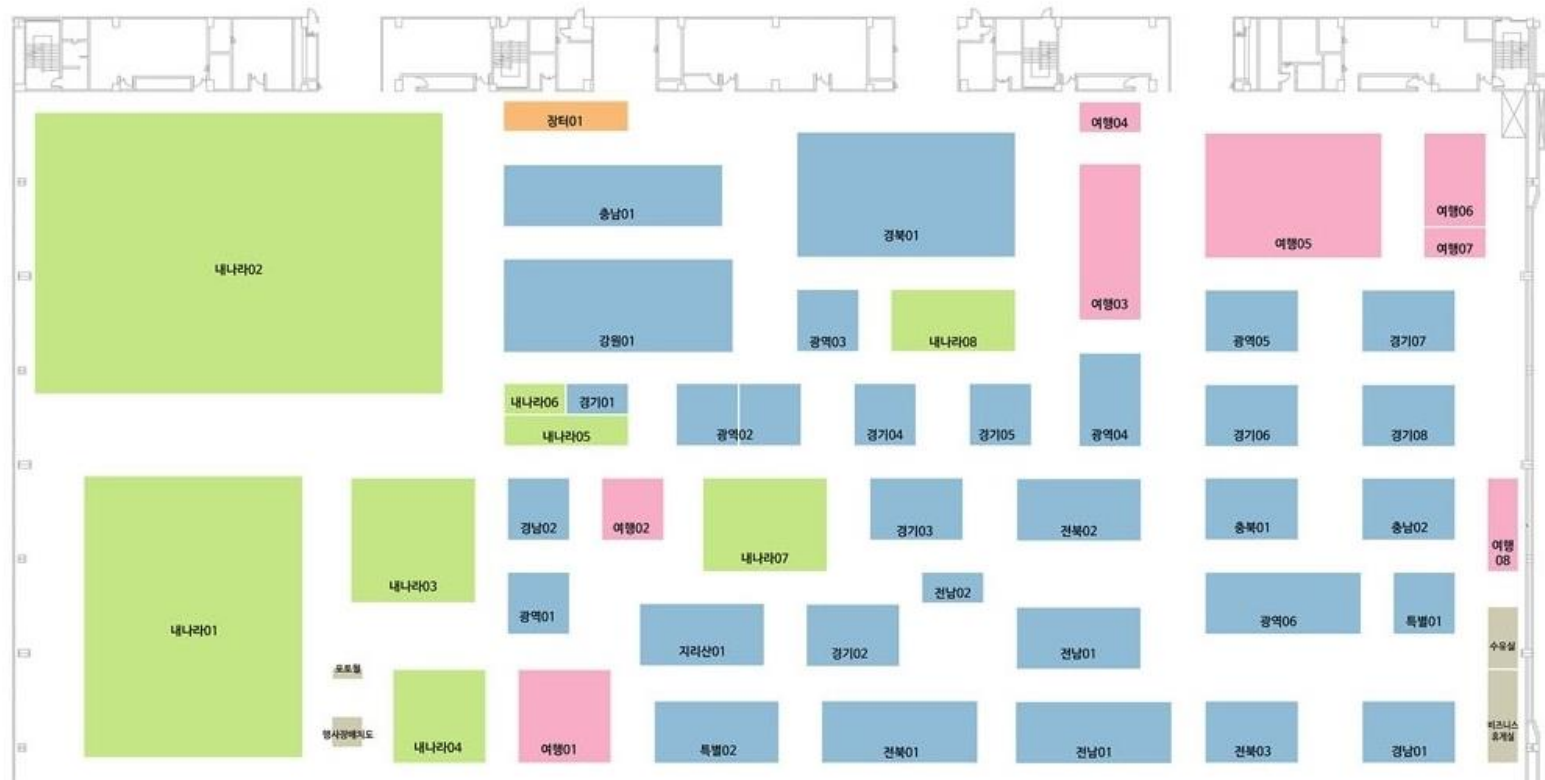
-  입찰공고서(2017 내나라 여행박람회 문화재청 전시부스 설치 운영 용역).hwp
-  과업내용서(2017 내나라 여행박람회 문화재청 전시부스 설치 운영용역).hwp

사례 3: 박람회 부스 이용 패턴 분석

❖ “2013 내나라 여행 박람회” 데이터를 이용한 분석 사례

- 연구 목적: 근접무선통신기술(NFC)을 이용하여 수집한 박람회 관람객 행동 데이터를 분석하여 전시부스 간 연관관계 분석 및 부스 배치 최적화

● 부스 배치도



사례 3: 박람회 부스 이용 패턴 분석

❖ “2013 내나라 여행 박람회” 데이터를 이용한 분석 사례

■ 연관규칙 분석 결과

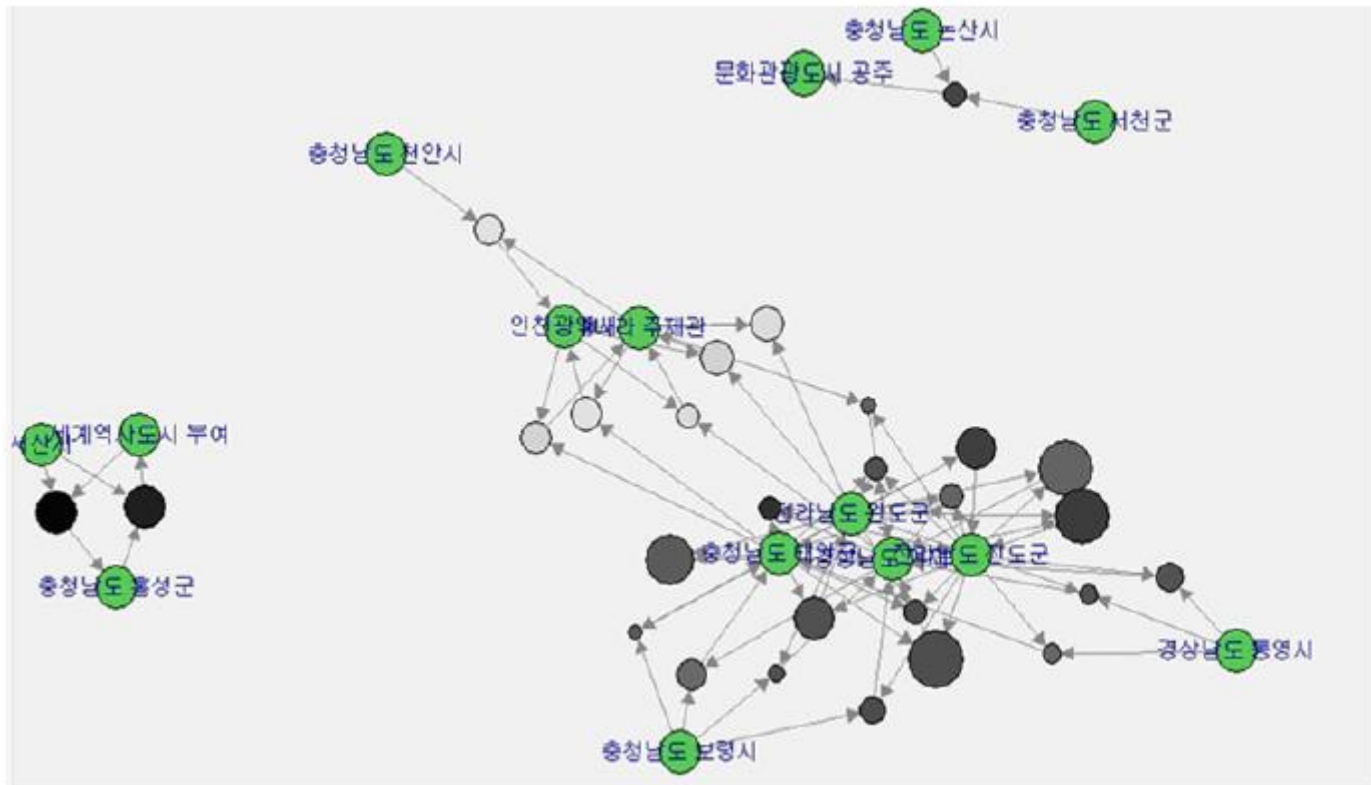
조건절	결과절	지지도	신뢰도	향상도
전남 진도, 충남 보령	경남 거제	0.098	0.889	4.040
경남 통영, 전남 진도	충남 태안	0.093	0.811	3.626
충남 논산, 충남 서천	충남 공주	0.096	0.850	4.145
인천광역시, 전남 완도	내나라 주제관	0.104	0.857	2.063

최명희, 전정호, 강희구, 이경전. (2013). 근접 무선 통신 기반 박람회 지원 시스템 구축 및 관람객 행동 데이터 분석 사례, Information Systems Review, 15(2): 111-127.

사례 3: 박람회 부스 이용 패턴 분석

❖ “2013 내나라 여행 박람회” 데이터를 이용한 분석 사례

■ 연관규칙 분석 결과



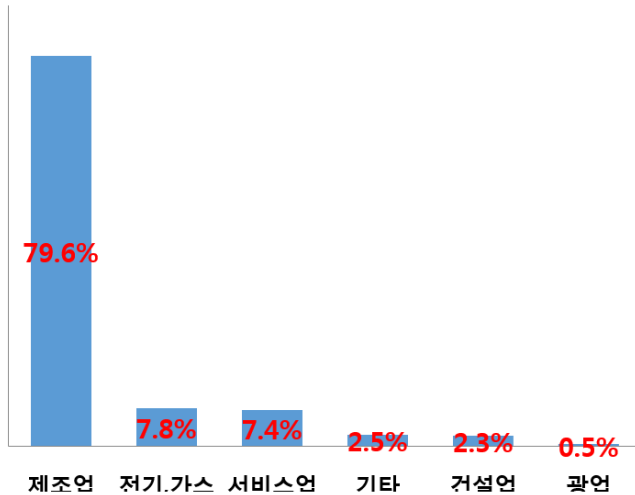
최명희, 전정호, 강희구, 이경전. (2013). 근접 무선 통신 기반 박람회 지원 시스템 구축 및 관람객 행동 데이터 분석 사례, *Information Systems Review*, 15(2): 111-127.

사례 4: 기업체 교육프로그램 추천

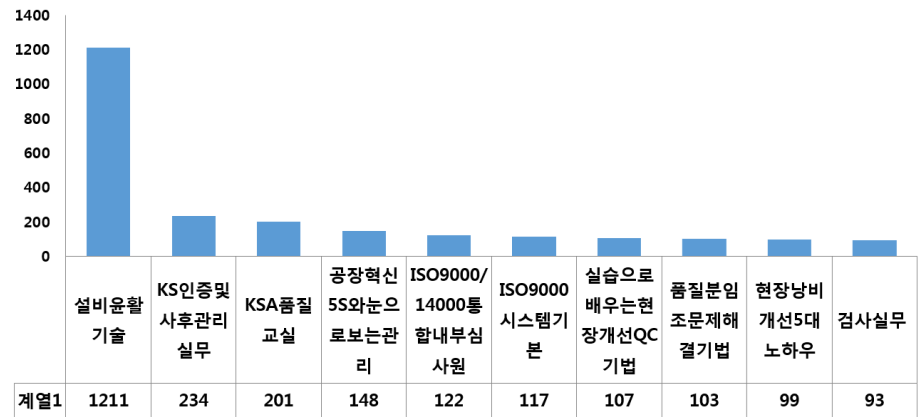
❖ 국내 K사의 기업체 교육 프로그램 운영 현황

- 교육 대상 기업의 주 업종은 제조업으로써 전체 교육 이수 기업의 79.6%에 해당함
- 대부분의 기업이 1 ~ 5개의 교육을 이수하고 있음

교육기관	기간	참여기업(수)	교육프로그램(수)	참가횟수
K사	2012 년 1월~8월	3,604	263	7,484



<참여기업의 업종별 분류>



<참여기업수 상위 10개 프로그램>

사례 4: 기업체 교육프로그램 추천

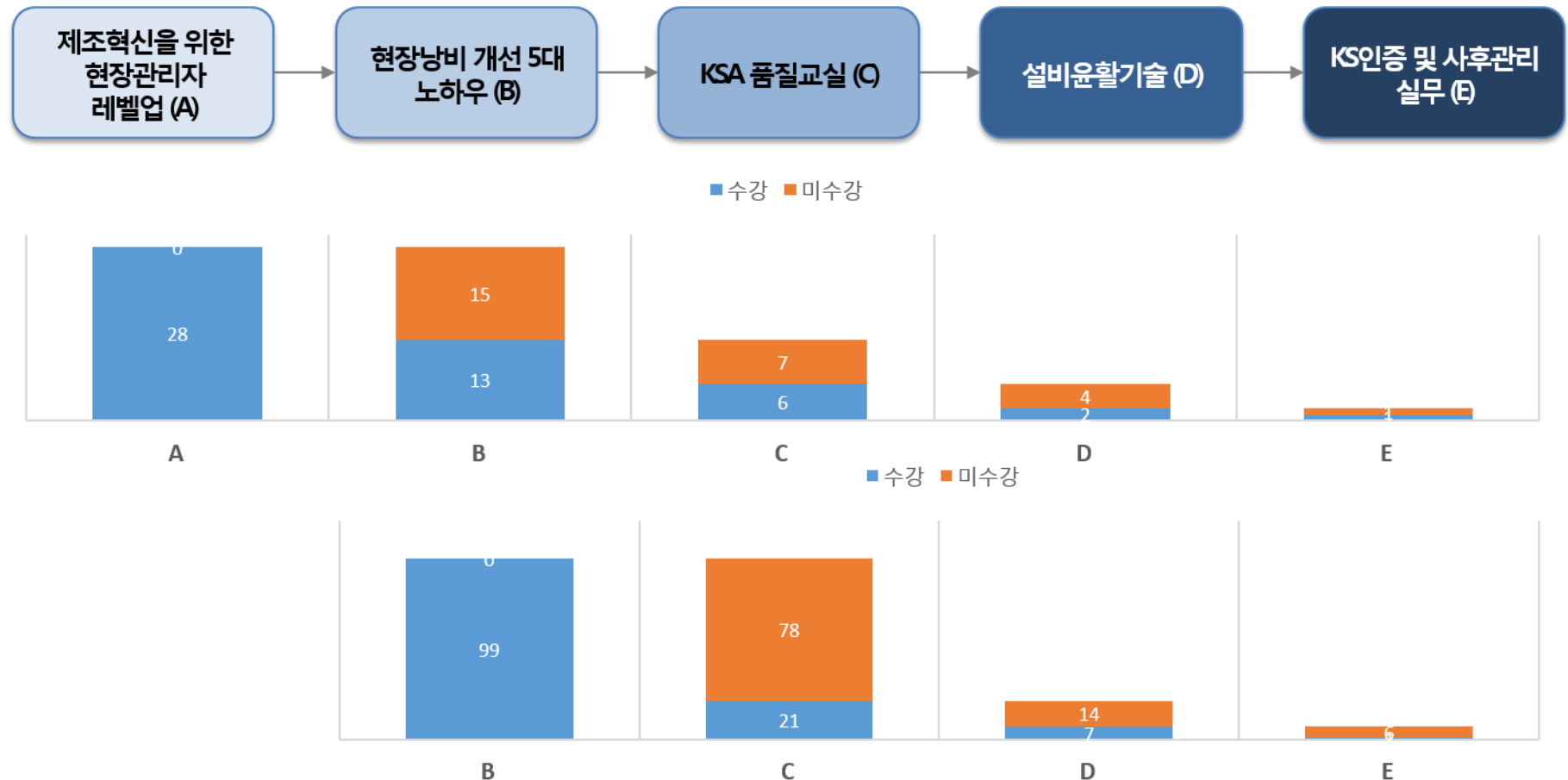
❖ 연관규칙분석 수행

- 지지도와 신뢰도의 기준값을 변화시켜 가며 총 25가지의 조합에 대해 생성된 규칙을 조사
- 현실적으로 활용 가능한 20~40개의 규칙을 생성한 세 가지 조합에 대한 추가 분석 실시
 - ✓ Case 1: 지지도 0.01, 신뢰도 0.2, Case 2: 지지도 0.01, 신뢰도 0.2, Case 3: 지지도 0.005, 신뢰도 0.3

		지지도(support)				
		0.002	0.005	0.01	0.015	0.02
신뢰도(confidence)	0.1	2,264	196	40	19	6
	0.2	1,363	106	24	12	4
	0.3	928	40	4	2	1
	0.4	688	16	3	1	1
	0.5	592	13	2	1	1

사례 4: 기업체 교육프로그램 추천

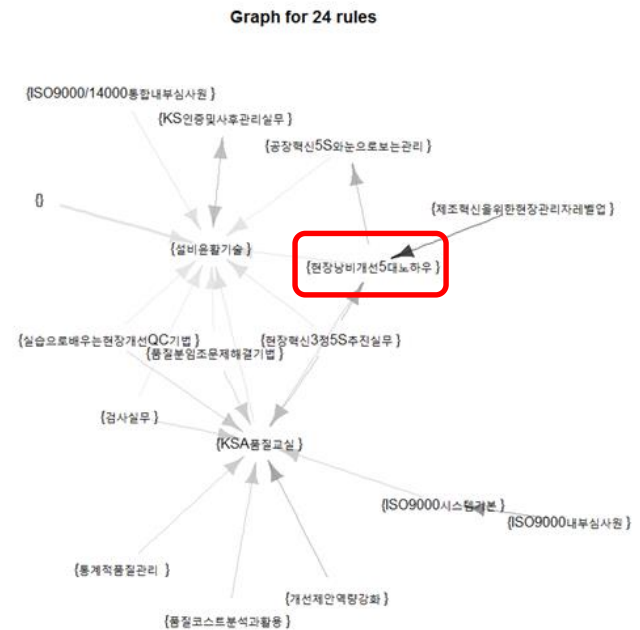
❖ 생성 규칙 분석 I: Chain 규칙 발굴을 통한 교육프로그램 단계적 추천



사례 4: 기업체 교육프로그램 추천

❖ 생성 규칙 분석 2: 수강기업수에 비해 규칙 등장 횟수가 높은 교육프로그램 발굴

순위	교육명	수강 기업수	순위	교육명	규칙등장 횟수
1	설비유효기술	1211	1	설비유효기술	11
2	KS인증및사후관리실무	234	2	KSA품질교실	10
3	KSA품질교실	201	3	현장낭비개선5대노하우	5
4	공장혁신5S와눈으로보는관리	148	4	현장혁신3정5S추진실무	3
5	ISO9000/14000통합내부심사원	122	5	품질분임조문제해결기법	2
6	ISO9000시스템기본	117	6	실습으로배우는현장개선QC기법	2
7	실습으로배우는현장개선QC기법	107	7	공장혁신5S와눈으로보는관리	2
8	품질분임조문제해결기법	103	8	검사실무	2
9	현장낭비개선5대노하우	99	9	KS인증및사후관리실무	2
10	검사실무	93	10	ISO9000시스템기본	2



사례 5: 한의학 처방전 분석

❖ 방약합편 분석

- 한의학 고문서에 나타난 텍스트를 분석하여 증상(symptoms)과 처방 약재(medicines) 간의 관계를 규명 (Yang et al., 2013)

연구대상 : 방약합편 (方藥合編)

- 한의학의 경험과 지식 축적은 주로 서적의 형태로 정리
- 방약합편 : 동의보감이 발간된 이후, 병증을 중심으로 두고 각 병증의 말미에 처방을 제시한 서적
- 병증에 따라 유용한 처방을 한눈에 제시
- 처방에 필요한 약재(본초)들을 명시



Materia medica/prescription	Ginseng Radix	Citri Pericarpium	Pinelliae Rhizoma	Poria(red)
Prescription 1	1	0	0	0		
Prescription 2	1	1	0	1		
Prescription 3	0	0	0	1		
Prescription 4	0	0	1	1		
...						
Prescription 521	1	1	0	0		

사례 5: 한의학 처방전 분석

❖ 방약합편 분석

- 한의학 고문서에 나타난 텍스트를 분석하여 증상(symptoms)과 처방 약재(medicines) 간의 관계를 규명 (Yang et al., 2013)

Antecedent (symptom)	Consequent (herbal materials)	Support (%)	Confidence (%)	Lift
Coughs	Pinelliae Rhizoma	4.2	46.0	2.2
	Citri Pericarpium	3.9	42.9	1.3
	Rehmanniae Radix Preparat	2.2	23.8	1.8
	Poria(red)	2.0	22.2	1.3
	Poria(white)	2.0	22.2	1.1
	Armeniacae Semen	1.9	20.6	3.0
Overexertion/Fatigue	Rehmanniae Radix Preparat	3.5	70.5	4.5
	Angelicae Gigantis Radix	3.3	67.6	1.9
	Ginseng Radix	2.0	41.1	1.4
	Poria(white)	1.9	38.2	1.9
	Dioscoreae Rhizoma	1.7	35.3	5.0
	Paeoniae Radix Alba	1.7	35.3	1.8
	Atractylodis Rhizoma Alba	1.6	32.3	1.2
	Corni Fructus	1.4	29.4	5.8
	Capreoli Cornu	1.3	26.5	8.2
	Astragali Radix	1.3	26.5	2.4
	Achyranthis Radix	1.2	23.5	5.0
	Pulvis Aconiti Tuberis Purificatum	1.2	23.5	2.8
	Cinnamomi Cortex Spissus	1.2	23.5	1.8
	Moutan Cortex	1.0	20.6	3.6
	Schizandrae Fructus	1.0	20.6	3.4

Table of Contents

I

연관규칙분석: A Priori Algorithm

II

연관규칙분석: 활용 분야

III

R 실습

R Exercise: 필요 패키지 설치 및 사용 준비

❖ 필요 패키지 설치 및 사용 준비

```
# Association Rules -----  
# arules and arulesViz packages install  
install.packages("arules", dependencies = TRUE)  
install.packages("arulesSequences", dependencies = TRUE)  
install.packages("arulesViz", dependencies = TRUE)  
install.packages("wordcloud", dependencies = TRUE)  
  
library(arules)  
library(arulesSequences)  
library(arulesViz)  
library(wordcloud)
```

- `install.packages()`: R 아카이브에 존재하는 패키지를 로컬 Machine으로 복사
 - ✓ 인터넷에 연결되어 있어야 하며 패키지가 업데이트되지 않는 이상 한번만 실행하면 됨
- `library()`: 로컬 Machine에 존재하는 package를 현재 스크립트에서 사용할 수 있도록 활성화

R Exercise: 필요 패키지 설치 및 사용 준비

❖ 필요 패키지 설치 및 사용 준비

```
# Association Rules -----  
# arules and arulesViz packages install  
install.packages("arules", dependencies = TRUE)  
install.packages("arulesSequences", dependencies = TRUE)  
install.packages("arulesViz", dependencies = TRUE)  
install.packages("wordcloud", dependencies = TRUE)  
  
library(arules)  
library(arulesSequences)  
library(arulesViz)  
library(wordcloud)
```

- arule: A-priori 알고리즘을 사용하여 연관규칙을 도출하는 기능 제공
- arulesSequences: SPADE 알고리즘을 사용하여 순차패턴을 효율적으로 탐색하는 기능 제공
- arulesViz: 연관규칙에 대한 시각화 기능 제공
- wordcloud: 단어 구름 도시 기능 제공

R Exercise: 데이터셋 불러오기

❖ 데이터셋 불러오기

```
# Part 1: Transform a data file into transaction format
# Basket type
tmp_basket <- read.transactions("Transaction_Sample_Basket.csv",
                                format = "basket", sep = ",", rm.duplicates=TRUE)
inspect(tmp_basket)
# Single type
tmp_single <- read.transactions("Transaction_Sample_Single.csv",
                                format = "single", cols = c(1,2), rm.duplicates=TRUE)
inspect(tmp_single)
```

- arules에서 제공하는 read.basket() 함수는 두 가지 형태의 데이터를 트랜잭션 데이터로 변환 가능
 - ✓ basket format: 행(row)이 트랜잭션 ID, 열(column)은 트랜잭션에 포함된 아이템
 - ✓ single format: 한 행은 트랜잭션 ID와 해당 트랜잭션에 포함된 1개의 아이템으로 구성

R Exercise: 데이터셋 불러오기

❖ 데이터셋 불러오기

- Basket format 형태 및 R에서 불러온 이후

	A	B	C	D	E
1	A	B	C		
2	A	C	D	E	
3	A	E	B		
4	B	C	D		
5	F	A	B		
6	A	D	F	G	
7	G	F	B	C	E
8	A	B			
9	C	D			
10	C	F	G		



```
> inspect(tmp_basket)
      items
[1] {A,B,C}
[2] {A,C,D,E}
[3] {A,B,E}
[4] {B,C,D}
[5] {A,B,F}
[6] {A,D,F,G}
[7] {B,C,E,F,G}
[8] {A,B}
[9] {C,D}
[10] {C,F,G}
```

R Exercise: 데이터셋 불러오기

❖ 데이터셋 불러오기

- Single format 형태 및 R에서 불러온 이후

	A	B
1	Tr1 A	
2	Tr1 B	
3	Tr1 C	
4	Tr2 A	
5	Tr2 C	
6	Tr2 D	
7	Tr2 E	
8	Tr3 A	
9	Tr3 E	
10	Tr3 B	
11	Tr4 B	
12	Tr4 C	
13	Tr4 D	
14	Tr5 F	
15	Tr5 A	
16	Tr5 B	
17	Tr6 A	
18	Tr6 D	
19	Tr6 F	
20	Tr6 G	



```
> inspect(tmp_single)
      items      transactionID
[1] {A,B,C}      Tr1
[2] {C,G}        Tr10
[3] {A,C,D,E}    Tr2
[4] {A,B,E}      Tr3
[5] {B,C,D}      Tr4
[6] {A,B,F}      Tr5
[7] {A,D,F,G}    Tr6
[8] {B,C,E,F,G}  Tr7
[9] {A,B}        Tr8
[10] {C}         Tr9
```

R Exercise I: 장바구니 데이터 분석

❖ 데이터 불러오기 및 확인

```
# Part 2: Association Rule Mining without sequence information
data("Groceries")
summary(Groceries)
str(Groceries)
inspect(Groceries)
```

- arules 패키지 설치 시 함께 제공되는 “Groceries” 데이터 사용

- ✓ Transaction 데이터의 형태이며 Sparse matrix로 저장되어 있음, 유용한 summary 정보를 함께 제공

```
> summary(Groceries)
transactions as itemMatrix in sparse format with
9835 rows (elements/itemsets/transactions) and
169 columns (items) and a density of 0.02609146

most frequent items:
      whole milk other vegetables      rolls/buns      soda      yogurt
      2513          1903          1809          1715          1372
      (Other)
      34055

element (itemset/transaction) length distribution:
sizes
  1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17
2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46   29
 18   19   20   21   22   23   24   26   27   28   29   32
 14   14    9   11    4    6    1    1    1    1    3    1

      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1.000   2.000   3.000   4.409   6.000  32.000

includes extended item information - examples:
      labels level2      level1
1 frankfurter sausage meat and sausage
2   sausage sausage meat and sausage
3   liver loaf sausage meat and sausage
~ |
```

R Exercise I: 장바구니 데이터 분석

❖ 구매한 아이템들을 이용한 Wordcloud 그리기

```
# Item inspection
itemName <- itemLabels(Groceries)
itemCount <- itemFrequency(Groceries)*9835
col <- brewer.pal(8, "Dark2")
wordcloud(words = itemName, freq = itemCount, min.freq = 1,
          scale = c(3, 0.2), col = col, random.order = FALSE)
```

- itemName: Wordcloud에 사용할 아이템 이름
- itemCount: Wordcloud에 사용할 아이템 빈도
- brewer.pal(): 사전에 정의된 색상 팔레트
- wordcloud(): Wordcloud 생성 함수
 - ✓ words: 사용 단어, freq: 사용 빈도, min.freq: 그래프 생성에 필요한 아이템 등장 최소 빈도,
scale: 최고빈도 단어와 최저빈도 단어 사이의 크기에 대한 상대적 비율,

```
# Item inspection
itemName <- itemLabels(Groceries)
itemCount <- itemFrequency(Groceries)*9835
col <- brewer.pal(8, "Dark2")
wordcloud(words = itemName, freq = itemCount, min.freq = 1,
          scale = c(3, 0.2), col = col, random.order = FALSE)
```

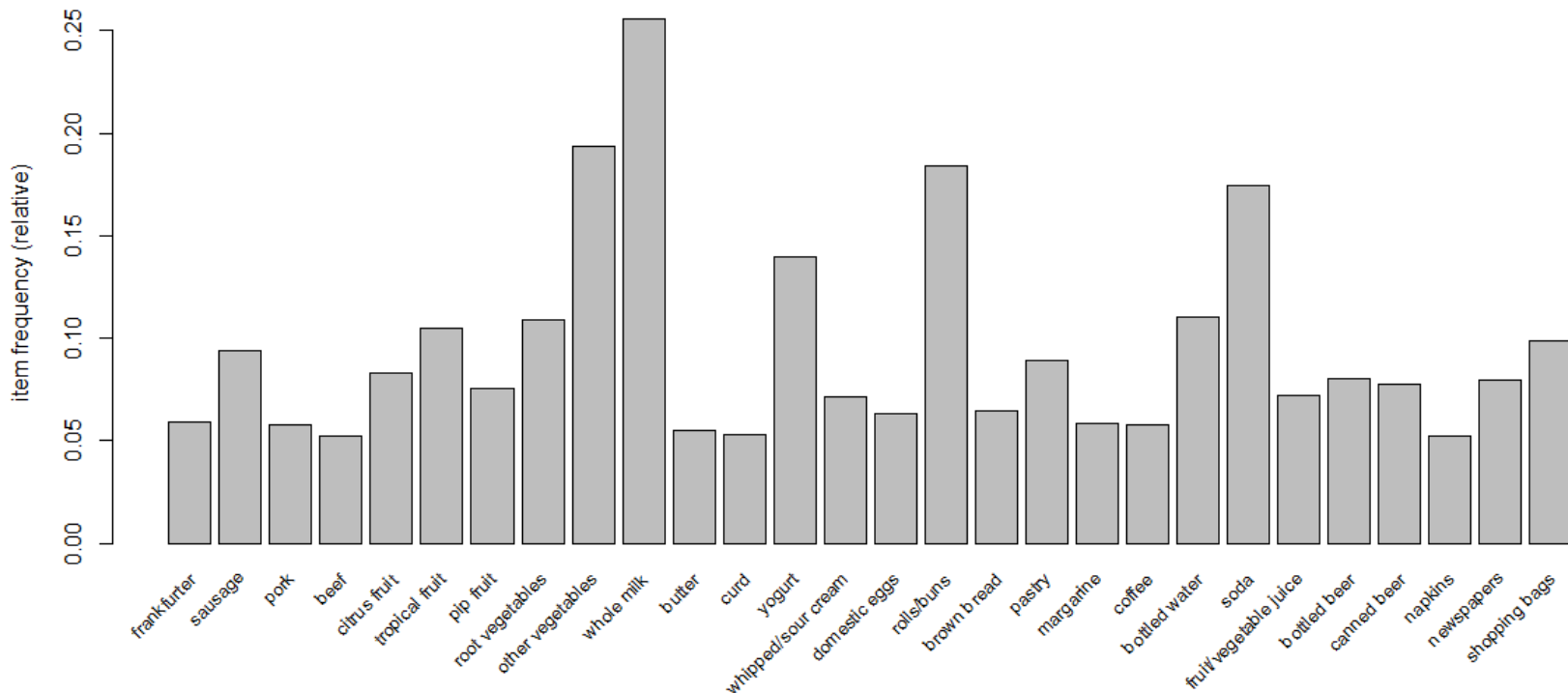


R Exercise I: 장바구니 데이터 분석

❖ 구매한 아이템별 빈도표 그리기

```
itemFrequencyPlot(Groceries, support = 0.05, cex.names=0.8)
```

- 최소 빈도 5% 이상 아이템들만 도시



R Exercise I: 장바구니 데이터 분석

❖ 규칙 생성

```
# Rule generation by Apriori
rules <- apriori(Groceries, parameter=list(support=0.01, confidence=0.35))

# Check the generated rules
inspect(rules)

# List the first three rules with the highest lift values
inspect(sort(rules, by="lift"))
```

- `inspect()` 내부의 명령어는 규칙을 lift 크기 순으로 내림차순 정렬해서 보여줄 것을 의미
- 총 89개의 규칙이 생성되었으며 lift 상위 10개 규칙은 아래와 같음

```
> inspect(sort(rules, by="lift"))
```

	lhs	rhs	support	confidence	lift	count
[1]	{citrus fruit,other vegetables}	=> {root vegetables}	0.01037112	0.3591549	3.295045	102
[2]	{citrus fruit,root vegetables}	=> {other vegetables}	0.01037112	0.5862069	3.029608	102
[3]	{tropical fruit,root vegetables}	=> {other vegetables}	0.01230300	0.5845411	3.020999	121
[4]	{whole milk,curd}	=> {yogurt}	0.01006609	0.3852140	2.761356	99
[5]	{root vegetables,rolls/buns}	=> {other vegetables}	0.01220132	0.5020921	2.594890	120
[6]	{root vegetables,yogurt}	=> {other vegetables}	0.01291307	0.5000000	2.584078	127
[7]	{tropical fruit,whole milk}	=> {yogurt}	0.01514997	0.3581731	2.567516	149
[8]	{yogurt,whipped/sour cream}	=> {other vegetables}	0.01016777	0.4901961	2.533410	100
[9]	{other vegetables,whipped/sour cream}	=> {yogurt}	0.01016777	0.3521127	2.524073	100
[10]	{root vegetables,whole milk}	=> {other vegetables}	0.02318251	0.4740125	2.449770	228

R Exercise I: 장바구니 데이터 분석

❖ 생성된 규칙 파일로 내보내기

```
# Save the rules in a text file
write.csv(as(rules, "data.frame"), "Groceries_rules.csv", row.names = FALSE)
```

- 생성된 규칙을 data.frame 형태로 변환한 뒤 csv파일 형식으로 저장
- ✓ xlsx 등의 MS Excel 파일 형식을 사용할수도 있으나 권장하지 않음 (속도 차이가 매우 크게 나타남)

	A	B	C	D	E	F
1	rules	support	confidence	lift	count	
2	{hard cheese} => {whole milk}	0.01006609	0.410788382	1.60768155	99	
3	{butter milk} => {other vegetables}	0.010371124	0.370909091	1.916915874	102	
4	{butter milk} => {whole milk}	0.011591256	0.414545455	1.622385414	114	
5	{ham} => {whole milk}	0.011489578	0.44140625	1.72750914	113	
6	{sliced cheese} => {whole milk}	0.010777834	0.439834025	1.721356003	106	
7	{oil} => {whole milk}	0.011286223	0.402173913	1.573967543	111	
8	{onions} => {other vegetables}	0.014234875	0.459016393	2.372268119	140	
9	{onions} => {whole milk}	0.012099644	0.390163934	1.526964702	119	
10	{berries} => {whole milk}	0.011794611	0.354740061	1.388328095	116	
11	{hamburger meat} => {other vegetables}	0.013828165	0.415902141	2.149446954	136	
12	{hamburger meat} => {whole milk}	0.014743264	0.443425076	1.735410118	145	
13	{hygiene articles} => {whole milk}	0.012811388	0.388888889	1.521974621	126	
14	{sugar} => {whole milk}	0.015048297	0.444444444	1.739399567	148	
15	{long life bakery product} => {whole milk}	0.013523132	0.361413043	1.414443805	133	
16	{dessert} => {whole milk}	0.013726487	0.369863014	1.447514023	135	
17	{cream cheese} => {whole milk}	0.016471784	0.415384615	1.625669595	162	
18	{chicken} => {other vegetables}	0.017895272	0.417061611	2.155439279	176	
19	{chicken} => {whole milk}	0.017590239	0.409952607	1.604410619	173	
20	{white bread} => {whole milk}	0.017081851	0.405797101	1.58814743	168	

R Exercise I: 장바구니 데이터 분석

❖ 생성된 규칙 그림으로 확인하기

```
# Plot the rules
plot(rules, method = "scatterplot")
plotly_arules(rules, method = "scatterplot", measure = c("support", "confidence"),
              shading = "lift")
```

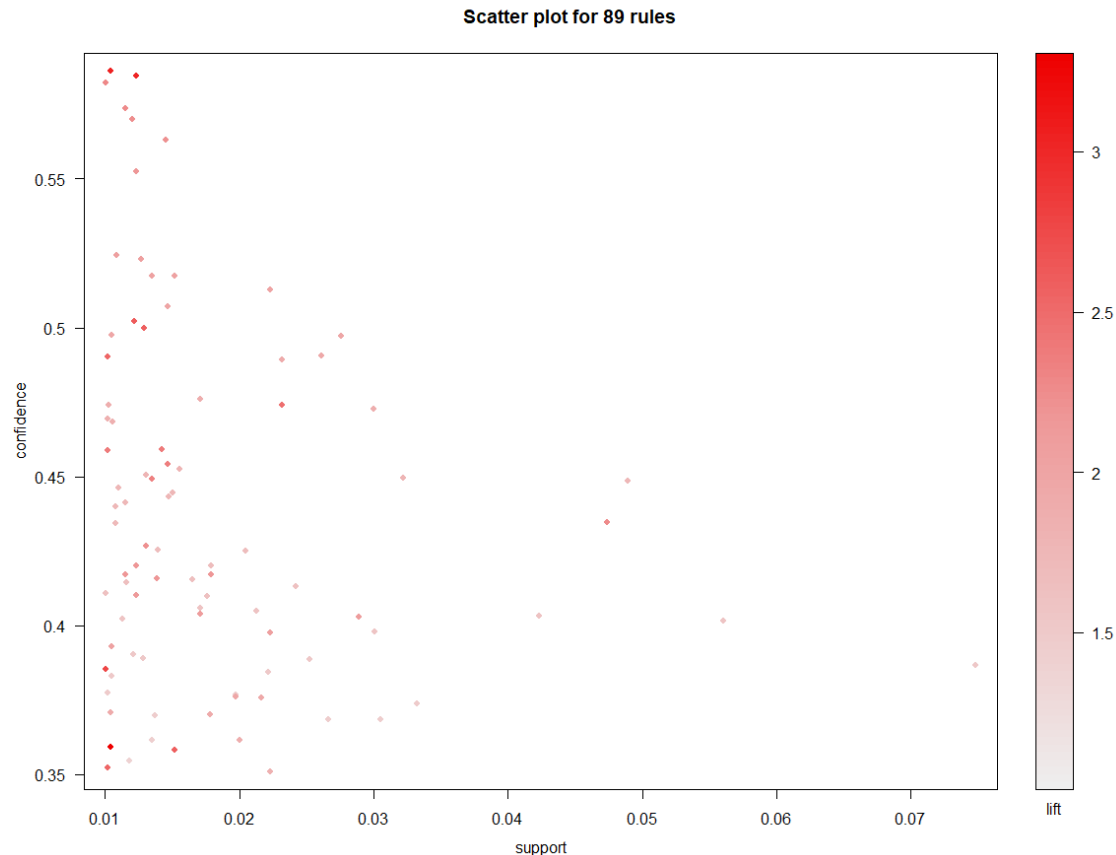
- `plot()` 함수는 생성된 규칙을 이용하여 고정된 그림을 그려주는 함수임
 - ✓ `method` 옵션에 `scatterplot`, `matrix`, `graph` 등의 다양한 형식을 지정할 수 있음
- `plotly_arules()` 함수는 사용자가 원하는 부분에 대한 조정이 가능한 `interactive plot`을 그려주는 기능을 제공함

R Exercise I: 장바구니 데이터 분석

❖ 생성된 규칙 그림으로 확인하기

■ `plot()` 함수 (`method = "scatterplot"`)

✓ 생성된 규칙의 지지도(x축), 향상도(y축), 신뢰도(색상)를 통한 데이터셋의 특징을 파악하는데 주로 사용

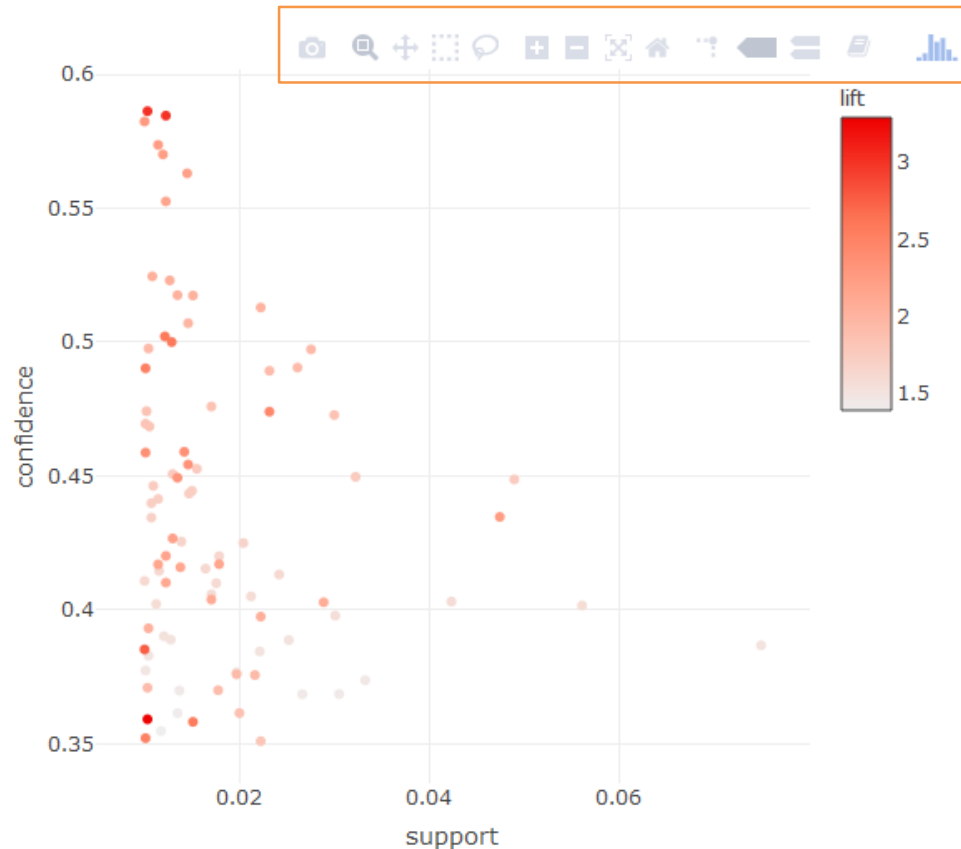


R Exercise I: 장바구니 데이터 분석

❖ 생성된 규칙 그림으로 확인하기

■ `plotly_arules()` 함수 (method = “scatterplot”)

✓ 기본 제공된 그림에서 사용자가 여러 옵션을 조정할 수 있음

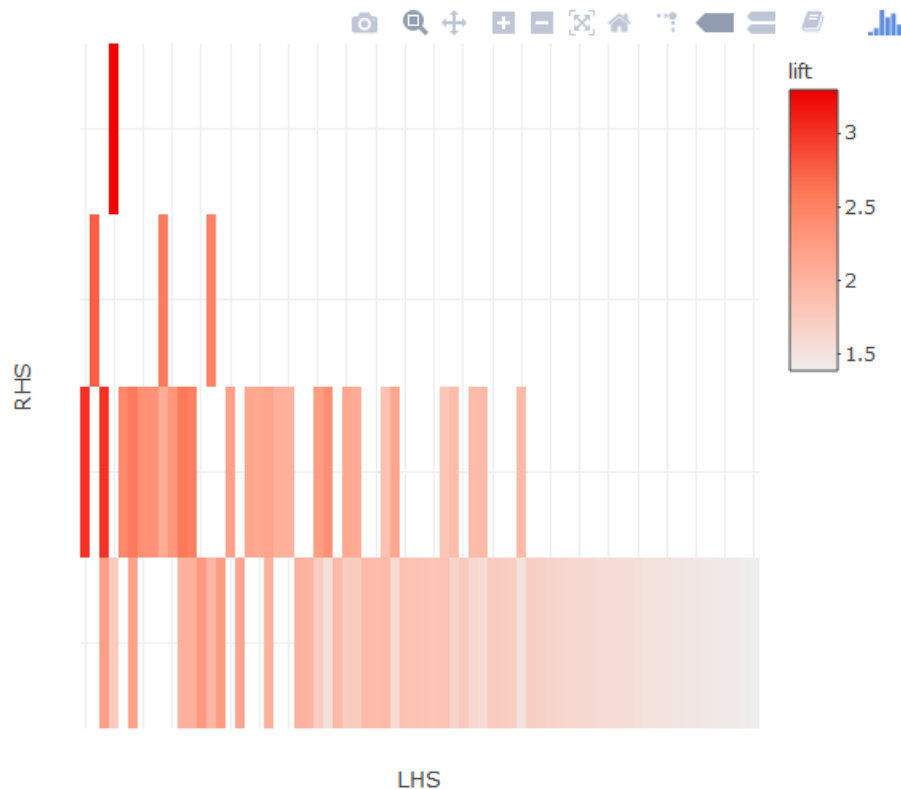


R Exercise I: 장바구니 데이터 분석

❖ 생성된 규칙 그림으로 확인하기

```
plot(rules, method="matrix")
plotly_arules(rules, method = "matrix", measure = c("support", "confidence"),
  shading = "lift")
```

- method = “matrix”를 사용한 경우: 조건절과 결과절 항목을 보다 명확히 표현



R Exercise I: 장바구니 데이터 분석

❖ 옵션을 바꾸어 보다 적은 수의 규칙 생성 및 도시

```
# Rule generation by Apriori with another parameters
rules <- apriori(Groceries, parameter=list(support=0.01, confidence=0.5))
plot(rules, method="graph")
plot(rules, method="paracoord")
```

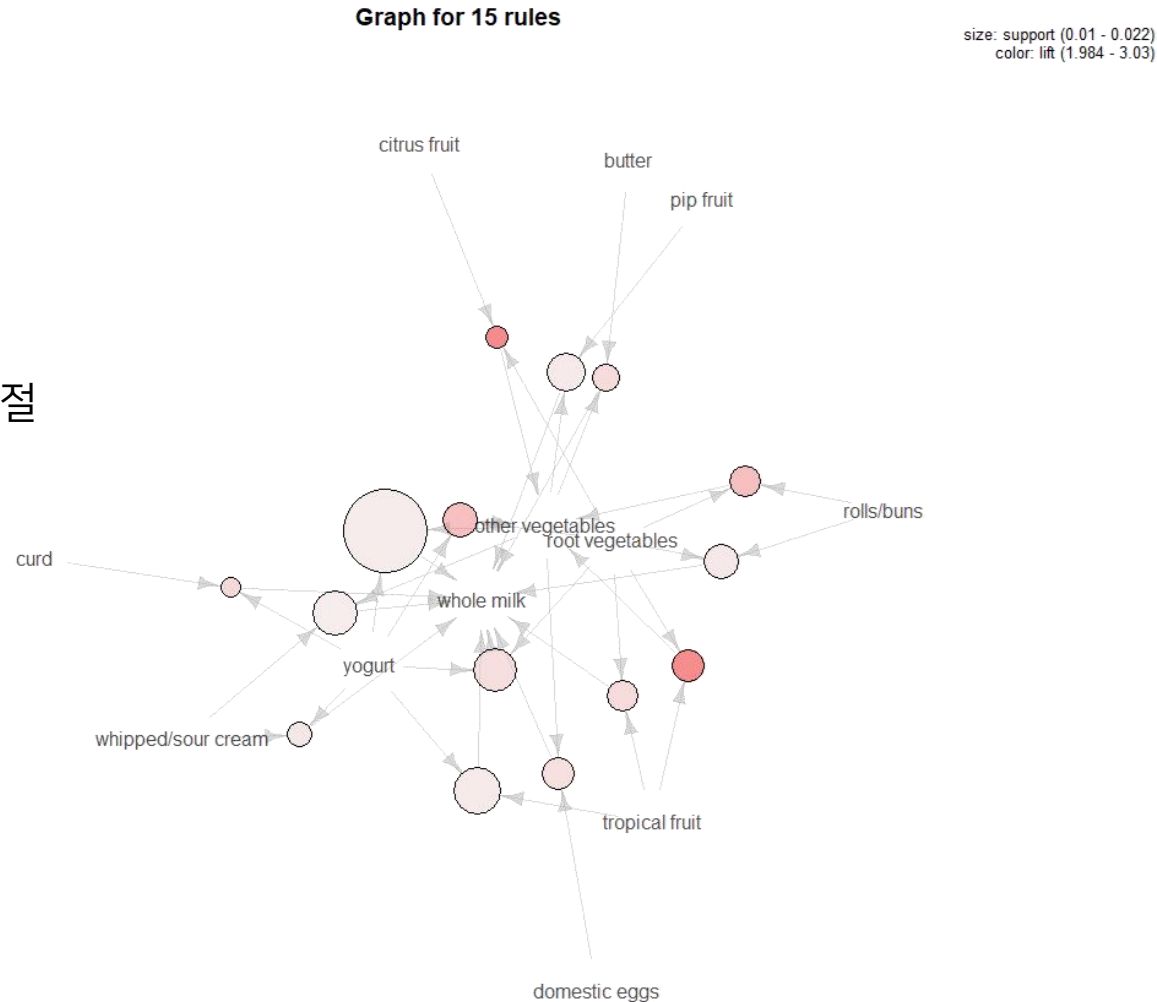
- confidence 기준을 0.35 → 0.5로 증가시킴
- “graph” method와 “paracoord” method는 규칙에서의 아이템간 관계에 보다 집중해서 그림을 그려주는 기능을 제공함

R Exercise I: 장바구니 데이터 분석

❖ 옵션을 바꾸어 보다 적은 수의 규칙 생성 및 도시

■ “graph” method 사용 시

- ✓ 원: 규칙
- ✓ 원의 크기: 지지도
- ✓ 원의 색상: 향상도
- ✓ 원으로 들어오는 화살표: 조건절
- ✓ 원에서 나가는 화살표: 결과절

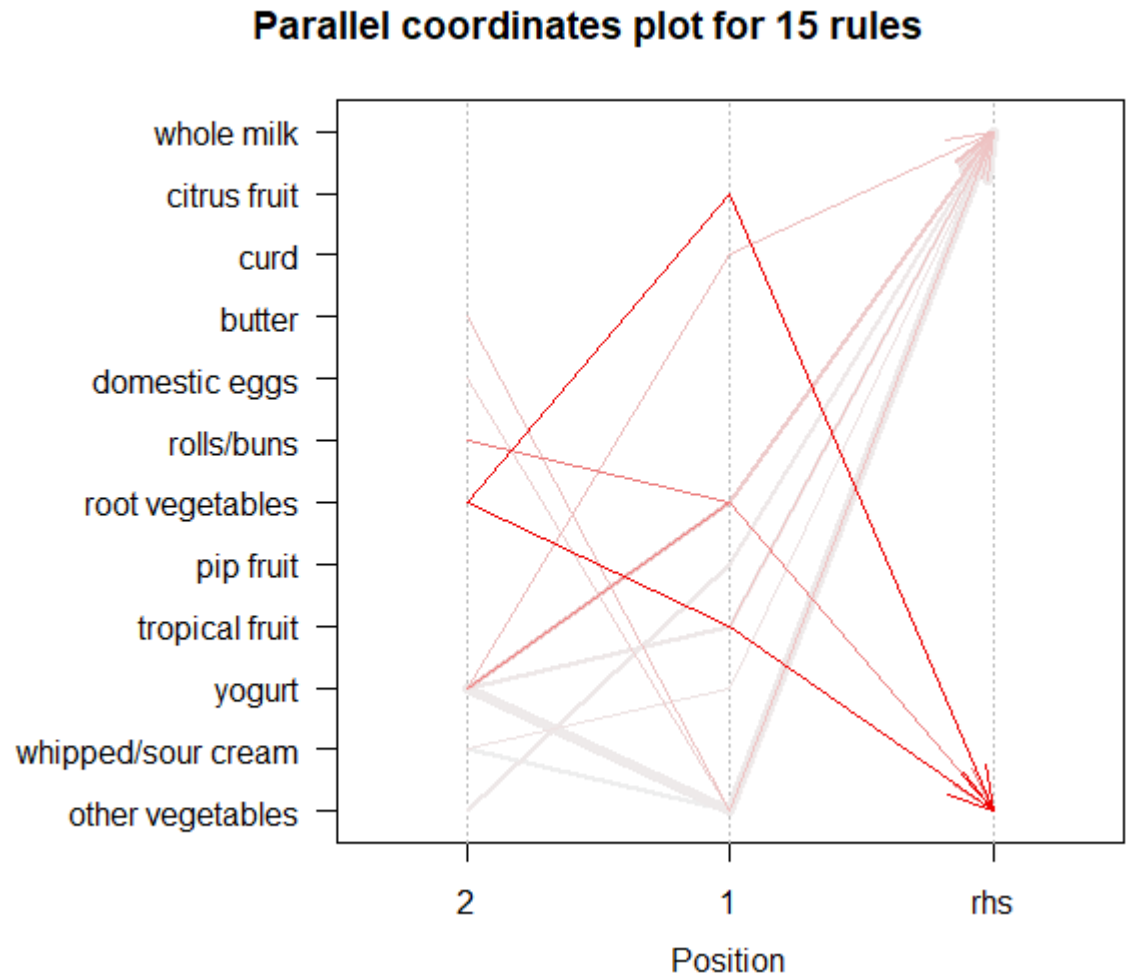


R Exercise I: 장바구니 데이터 분석

❖ 옵션을 바꾸어 보다 적은 수의 규칙 생성 및 도시

▪ “paracoord” method 사용 시

- ✓ 선: 규칙
- ✓ x축: 아이템 순서
- ✓ y축: 해당 아이템 명



R Exercise 2: 순차연관분석

❖ 데이터셋

```
# Part 3: Association Rule Mining with sequence information
foodmart_tr <- read_baskets("foodmart_transactions2.txt",
  info = c("sequenceID", "eventID", "SIZE"))
foodmart_df <- as(foodmart_tr, "data.frame")
```

- Foodmart transaction data: MS SQL 2000에서 기본으로 제공하는 Foodmart의 데이터베이스
 - ✓ 이 중 고객 정보와 거래 정보를 가진 세 가지의 테이블을 사용하여 연관규칙분석에서 사용할 수 있는 형태로 미리 변환
 - ✓ sequenceID: 고객 ID
 - ✓ eventID: 한 고객이 구매한 시점에 대한 ID
 - ✓ SIZE: 해당 트랜잭션에서 구매한 아이템의 총 수

R Exercise 2: 순차연관분석

❖ 데이터셋 예시

Filter				
	items	sequenceID	eventID	SIZE
1	{CDR_Hot_Chocolate,Faux_Products_Laundry_Detergent,High_Top_Beets,Lake_Low_Fat_Cole_Slaw,Monarch_Spaghetti,Skinner_Strawberry_Drink,Super_Brown_Sugar}	3	117	7
2	{American_Roasted_Chicken,Big_Time_Low_Fat_Waffles,Consolidated_Tartar_Control-Toothpaste,Imagine_Chicken_TV_Dinner,Pleasant_Canned_Tuna_in_Water,Red_Wing_60_Wa...	3	295	6
3	{Carrington_Frozen_Chicken_Thighs,High_Top_Mandarin_Oranges,Hilltop_Laundry_Detergent,Just_Right_Large_Canned_Shrimp}	3	331	4
4	{Bird_Call_200_MG_Acetominifen,Bird_Call_Conditioning_Shampoo,Ebony_Lettuce,Even_Better_String_Cheese,Gauss_Monthly_Home_Magazine,Gorilla_Havarti_Cheese,Tell_Tale_...	3	453	7
5	{Consolidated_200_MG_Acetominifen,Gorilla_Chocolate_Milk,Gorilla_String_Cheese,Monarch_Thai_Rice,Musial_Spicy_Mints,Red_Wing_Copper_Pot_Scrubber}	3	472	6
6	{Moms_Sliced_Turkey}	5	4	1
7	{Atomic_Bubble_Gum,BBB_Best_Strawberry_Preserves,Club_Sharp_Cheddar_Cheese,Gauss_Monthly_Computer_Magazine,Just_Right_Canned_Yams,Modell_Cranberry_Muffins}	6	204	6
8	{BBB_Best_Regular_Coffee,Red_Spade_Potato_Salad}	6	381	2
9	{Bravo_Chicken_Noodle_Soup,Even_Better_1%_Milk,Gulf_Coast_Bubble_Gum,Urban_Large_Eggs}	6	541	4
10	{Booker_Havarti_Cheese,Horatio_Dried_Apricots,Imagine_Frozen_Chicken_Breast,Sunset_Plastic_Forks}	8	426	4

R Exercise 2: 순차연관분석

❖ 빈발 순차 패턴 추출

```
# Find frequent sequences
start.time <- proc.time()
seq_rules <- cspade(foodmart_tr, parameter = list(support = 0.0005,
          maxsize = 10, maxlen = 5), control = list(verbose = TRUE))
proc.time() - start.time

summary(seq_rules)
```

- `proc.time()`: 현재 시간을 기록하는 함수
- `cspade()`: 빈발 순차 패턴을 추출하는 함수
 - ✓ `support`: 빈발 순차 패턴의 최소 지지도
 - ✓ `maxsize`: 한 구매(event)에 포함될 수 있는 최대 아이템 수
 - ✓ `maxlen`: 빈발 순차 패턴의 최대 길이(최대 이벤트 수)
- 참고: i7-7500U CPU 사용 Laptop에서 51.52초 소요

R Exercise 2: 순차연관분석

❖ 빈발 순차 패턴 추출

```
> summary(seq_rules)
set of 463805 sequences with

most frequent items:
      Better_Canned_Tuna_in_Oil      Steady_Childrens_Cold_Remed
      1177                        1174
Hilltop_Silky_Smooth_Hair_Conditioner  Ebony_Mixed_Nuts
      1123                        1090
      Hermanos_Golden_Delcious_Apples
      1166
      (Other)
      919948

most frequent elements:
      {Better_Canned_Tuna_in_Oil}      {Steady_Childrens_Cold_Remed
      1156                        1151
      {Hilltop_Silky_Smooth_Hair_Conditioner}  {Skinner_Strawberry_Drink
      1091                        1071
      {Hermanos_Golden_Delcious_Apples}
      1141
      (Other)
      909345

element (sequence) size distribution:
sizes
      1      2      3
11915 451855      35

sequence length distribution:
lengths
      1      2      3      4      5      6
1559 461901      299      36      9      1
```

- most frequent items: 빈발 순차 패턴에 빈번하게 등장하는 개별 아이템
- most frequent elements: 빈발 순차 패턴에 빈번하게 등장하는 아이템셋

R Exercise 2: 순차연관분석

❖ 빈발 순차 패턴 추출

```
# Filter frequent sequences with the length greater than 2
seq_rules_df <- as(seq_rules, "data.frame")
seq_rules_size <- size(seq_rules)
seq_rules_df <- cbind(seq_rules_df, seq_rules_size)
seq_rules_df_filtered <- subset(seq_rules_df, seq_rules_df$seq_rules_size > 2)
write.csv(seq_rules_df_filtered, file = "seq_rules_filterd.csv", row.names = FALSE)
```

- line 1: 생성된 빈발 순차 패턴을 데이터프레임으로 저장
- line 2: 생성된 빈발 순차 패턴의 길이 (eventID의 개수)를 저장
- line 3: line 1의 결과물과 line 2의 결과물을 결합
- line 4: line 3의 결과물에서 빈발 순차 패턴의 길이가 2 초과인 패턴들만 저장
- line 5: line 4의 결과물을 csv 형태로 저장

R Exercise 2: 순차연관분석

❖ 빈발 순차 패턴 추출

■ 추출된 빈발 순차 패턴

	A	B	C	D
1	sequence	support	seq_rules_size	
2	<{Fast_Golden_Raisins},{Fast_Frosted_Donuts},{Walrus_Imported_Beer}>	0.000572	3	
3	<{Curlew_Lox},{Red_Spade_Sliced_Chicken},{Thresher_Mint_Chocolate_Bar}>	0.000572	3	
4	<{Red_Wing_Scented_Toilet_Tissue},{Tell_Tale_Squash},{Super_Grape_Jelly}>	0.000572	3	
5	<{Just_Right_Regular_Ramen_Soup},{Big_Time_Frozen_Chicken_Wings},{Super_Apple_Jam}>	0.000572	3	
6	<{Great_English_Muffins},{Symphony_Rosy_Sunglasses},{Steady_Childrens_Cold_Remedy}>	0.000572	3	
7	<{Even_Better_Whole_Milk},{Even_Better_Jack_Cheese},{Ship_Shape_Seasoned_Hamburger}>	0.000687	3	
8	<{Hermanos_Lettuce},{Even_Better_Large_Curd_Cottage_Cheese},{Red_Wing_Bees_Wax_Candles}>	0.000572	3	
9	<{Cormorant_Scissors},{Nationeel_Sugar_Cookies},{Red_Wing_75_Watt_Lightbulb}>	0.000572	3	
10	<{High_Top_New_Potatos},{Akron_City_Map},{Plato_Low_Fat_Apple_Butter}>	0.000572	3	
11	<{Skinner_Mango_Drink},{Carrington_Waffles},{Plato_Extra_Chunky_Peanut_Butter}>	0.000572	3	
12	<{Sunset_AA-Size_Batteries},{Robust_Monthly_Computer_Magazine},{Nationeel_Beef_Jerky}>	0.000572	3	
13	<{Nationeel_Potato_Chips},{Cormorant_Paper_Cups},{Monarch_Rice_Medly}>	0.000572	3	
14	<{Big_Time_Frozen_Mushroom_Pizza},{Choice_Spicy_Mints},{Moms_Potato_Salad}>	0.000572	3	
15	<{Hermanos_Golden_Delicious_Apples},{Pleasant_Regular_Ramen_Soup},{Just_Right_Chicken_Noodle_Soup}>	0.000572	3	
16	<{Black_Tie_Eyeglass_Screwdriver},{Better_Canned_Beets},{Just_Right_Chicken_Noodle_Soup}>	0.000572	3	
17	<{Plato_Strawberry_Jam},{Skinner_Strawberry_Drink},{Imagine_Grape_Popsicles}>	0.000572	3	
18	<{Fort_West_Frosted_Donuts},{Blue_Label_Canned_Beets},{Gulf_Coast_Malted_Milk_Balls}>	0.000572	3	
19	<{Fort_West_Avocado_Dip},{Bravo_Noodle_Soup},{Great_Muffins}>	0.000572	3	
20	<{Horatio_Fondue_Mix},{Great_Rye_Bread},{Fast_Potato_Chips}>	0.000572	3	

R Exercise 2: 순차연관분석

❖ 순차 연관규칙 추출

```
# Find association rules
seq_rules_induced <- ruleInduction(seq_rules, confidence = 0.2)
summary(seq_rules_induced)
inspect(seq_rules_induced)

# Save the results
seq_rules_induced_df <- as(seq_rules_induced, "data.frame")
write.csv(seq_rules_induced_df, file = "seq_rules_induced_df.csv",
          row.names = FALSE)
```

- ruleInduction()함수: 빈발 순차패턴에서 최소 confidence 기준을 만족하는 연관규칙분석 탐색

R Exercise 2: 순차연관분석

❖ 순차 연관규칙 추출

■ 추출 결과

- ✓ 총 35개의 규칙 생성
- ✓ rule size distribution:
조건절과 결과절에 포함된
개별 아이템의 수는 모두 3개
- ✓ rule length distribution:
조건절과 결과절에 포함된
events의 수는 모두 3개
- ✓ 즉 한 event에 하나의 item만
속하 3개 순차 event들만이
연관규칙으로 도출됨

```
> summary(seq_rules_induced)
set of 35 sequencerules with

rule size distribution (lhs + rhs)
sizes
 3
35

rule length distribution (lhs + rhs)
lengths
 3
35

summary of quality measures:
      support      confidence      lift
Min.   :0.0005723  Min.   :0.4167  Min.   :22.56
1st Qu.:0.0005723  1st Qu.:0.5556  1st Qu.:31.63
Median :0.0005723  Median :0.7143  Median :37.14
Mean   :0.0005756  Mean   :0.6858  Mean   :37.09
3rd Qu.:0.0005723  3rd Qu.:0.7143  3rd Qu.:39.98
Max.   :0.0006868  Max.   :1.0000  Max.   :59.84

mining info:
      data ntransactions nsequences support confidence
foodmart_tr      54537      8736 5e-04      0.2
```

R Exercise 2: 순차연관분석

❖ 순차 연관규칙 추출

■ 추출 결과

	A	B	C	D
rule	support	confidence	lift	
<{Fast_Golden_Raisins},{Fast_Frosted_Donuts}> => <{Walrus_Imported_Beer}>	0.000572	0.8333333333	48.21192	
<{Curlew_Lox},{Red_Spade_Sliced_Chicken}> => <{Thresher_Mint_Chocolate_Bar}>	0.000572	0.5	25.84615	
<{Red_Wing_Scented_Toilet_Tissue},{Tell_Tale_Squash}> => <{Super_Grape_Jelly}>	0.000572	0.714285714	39.49367	
<{Just_Right_Regular_Ramen_Soup},{Big_Time_Frozen_Chicken_Wings}> => <{Super_Apple_Jam}>	0.000572	0.5	26	
<{Great_English_Muffins},{Symphony_Rosy_Sunglasses}> => <{Steady_Childrens_Cold_Remedy}>	0.000572	1	44.57143	
<{Even_Better_Whole_Milk},{Even_Better_Jack_Cheese}> => <{Ship_Shape_Seasoned_Hamburger}>	0.000687	0.545454545	29.59684	
<{Hermanos_Lettuce},{Even_Better_Large_Curd_Cottage_Cheese}> => <{Red_Wing_Bees_Wax_Candles}>	0.000572	0.714285714	39.24528	
<{Cormorant_Scissors},{Nationeel_Sugar_Cookies}> => <{Red_Wing_75_Watt_Lightbulb}>	0.000572	0.714285714	35.45455	
<{High_Top_New_Potatos},{Akron_City_Map}> => <{Plato_Low_Fat_Apple_Butter}>	0.000572	0.8333333333	44.12121	
<{Skinner_Mango_Drink},{Carrington_Waffles}> => <{Plato_Extra_Chunky_Peanut_Butter}>	0.000572	0.714285714	39.49367	
<{Sunset_AA-Size_Batteries},{Robust_Monthly_Computer_Magazine}> => <{Nationeel_Beef_Jerky}>	0.000572	0.5555555556	31.11111	
<{Nationeel_Potato_Chips},{Cormorant_Paper_Cups}> => <{Monarch_Rice_Medly}>	0.000572	0.625	30	
<{Big_Time_Frozen_Mushroom_Pizza},{Choice_Spicy_Mints}> => <{Moms_Potato_Salad}>	0.000572	0.8333333333	40.22099	
<{Hermanos_Golden_Delicious_Apples},{Pleasant_Regular_Ramen_Soup}> => <{Just_Right_Chicken_Noodle_Soup}>	0.000572	0.5	26.79755	
<{Black_Tie_Eyeglass_Screwdriver},{Better_Canned_Beets}> => <{Just_Right_Chicken_Noodle_Soup}>	0.000572	0.8333333333	44.66258	
<{Plato_Strawberry_Jam},{Skinner_Strawberry_Drink}> => <{Imagine_Grape_Popsicles}>	0.000572	0.625	38.4507	
<{Fort_West_Frosted_Donuts},{Blue_Label_Canned_Beets}> => <{Gulf_Coast_Malted_Milk_Balls}>	0.000572	0.714285714	38.28221	
<{Fort_West_Avocado_Dip},{Bravo_Noodle_Soup}> => <{Great_Muffins}>	0.000572	0.714285714	41.3245	
<{Horatio_Fondue_Mix},{Great_Rye_Bread}> => <{Fast_Potato_Chips}>	0.000572	1	59.02703	

Q & A

